Proceedings of the Prosody-Discourse Interface Conference 2013 (IDP-2013)

Edited by Piet Mertens
& Anne Catherine Simon

ISBN: 9-789090278-76-6

Conference website: http://wwwling.arts.kuleuven.be/franitalco/idp2013

Sponsors

FranItalCo

Institut Langage & Communication

Scientific committee

- Corine Astésano, Université de Toulouse II - Le Mirail
- Antoine Auchlin, Université de Genève
- Roxane Bertrand, LPL, Université Aix-Marseille
- Nicole Dehé, Universität Konstanz
- Elisabeth Delais-Roussarie, LLF, Université Paris-Diderot
- Céline De Looze, Trinity College Dublin
- Elwys de Stefani, University of Leuven (KU Leuven)
- Daniel Hirst, LPL, Université Aix-Marseille
- Mariapaola d'Imperio, LPL, Université Aix-Marseille
- Anne Grobet, Université de Genève, Ecole de langue et de civilisation française
- Anne Lacheret, MoDyCo, Université Paris Ouest
- Philippe Martin, Clillac-Arp, Université Paris-Diderot
- Piet Mertens, University of Leuven (KU Leuven)
- Anne Catherine Simon, Université catholique de Louvain
- Cristel Portes, LPL, Université Aix-Marseille
- Brechtje Post, University of Cambridge
- Petra Wagner, Universität Bielefeld
- Anne Wichmann, University of Central Lancashire

Local organizing committee

Convenors

- Piet Mertens, University of Leuven (KU Leuven)
- Anne Catherine Simon, Université catholique de Louvain

Collaborators

- Alice Bardiaux, Université catholique de Louvain
- Laurence Martin, Université catholique de Louvain
- Gélase Nimbona, Université catholique de Louvain
- Tom Velghe, University of Leuven (KU Leuven)

How to cite?

Mertens, Piet & Anne Catherine Simon (eds). 2013. *Proceedings of the Prosody-Discourse Interface Conference 2013 (IDP-2013).* Leuven, September 11-13, 2013. ISBN: 9-789090278-76-6, http://wwwling.arts.kuleuven.be/franitalco/idp2013/Proceedings.html

Bardiaux, Alice. (2013). Indices prosodiques régionaux en français de Belgique. L'apport d'une catégorisation perceptive des données, in Mertens, P. & A.C. Simon (eds), *Proceedings of the Prosody-Discourse Interface Conference 2013 (IDP-2013).* Leuven, September 11-13, 2013, pp. 13-19, http://wwwling.arts.kuleuven.be/franitalco/idp2013/Proceedings.html

Table of contents

Keynote speeches

Regular papers

Invited conference

Petra Wagner
University of Bielefeld

Timing and Prominence in Conversational Interaction

Conversations consist of interlocutors speaking in consecutive turns, but also show instances of simultaneous speech. e.g. when listeners actively provide feedback, e.g. in the form of backchannels not occuring during pauses. This dynamic timing in interaction needs a logistic component of utterance management which seems to be linked to the interlocutors' prosodic structure, i.e. the timing of syllabic units and prominent events initiating prosodic feet (Wlodarczak et al. 2011, 2012, in press). This logistic component, which we call *Interaction Phonology*, must therefore rely on the rhythmic phonological structure of individual languages (Wagner et al., in press). In our approach, we furthermore argue that this form of inter-speaker adaptation or *entrainment* enables interlocutors to guide their attention to relevant phonetic detail and to attune to the fine-grained organization underlying the linguistic structure encoded in the incoming speech signal. Entrainment processes can be formally modelled by dynamically phase and period adapting oscillators (Malisz et al., 2012; Inden et al., 2012; Wagner et al., 2012). Our assumptions were tested empirically by comparing various approaches to modeling listener feedback productions of an artificial agent (Inden et al., submitted).

As our model relies on identification of prominent events in the interlocutor‘s speech, an identification of these prominences needs to be modelled as well. While there seems to be a lot of agreement about the acoustic factors constituting prominent events across languages (quality and excursion of f0 movement, duration, spectral intensity), the language specificity of these "ingredients" is not well understood. Besides, the interaction with other potentially prominence lending cues is not well understood, among them being glottalization or creak, aspiration, consonantal lengthening and intensity and multimodal cues. Another aspect of prominence in conversations is the distinction between relative prominence and absolute prominence: Relative prominences express classic phonological functions such as lexical and phrasal stress or prosodic focus, while absolute prominence conveys a speaker‘s involvement, level of attention or pragmatic functions such as urgency of a message conveyed (Wagner and Portele, 1999).

In my talk, I will make suggestions as to how these complex interactions can be studied and modeled within and across languages and speaking styles (Wagner et al., 2012a; Wagner et al., 2012b; Arnold et al., 2011a,b; Arnold et al., 2012).

Arnold, D., Wagner, P., & Möbius, B. (2011). Evaluating different rating scales for obtaining judgments of syllable prominence from naïve listeners. International Congress of the Phonetic Sciences (pp. 252–255), Hong Kong, China.

Arnold, D., Möbius, B., & Wagner, P. (2011). Comparing word and syllable prominence rated by naive listeners. Proceedings of Interspeech 2011 (pp. 1877–1880), Florence, Italy.

Arnold, D., Wagner, P., & Möbius, B. (2012). Obtaining prominence judgments from naïve listeners – Influence of rating scales, linguistic levels and normalisation. *Proceedings of Interspeech 2012*, Portland, USA.

Inden, B., Malisz, Z., Wagner, P., & Wachsmuth, I. (2012). Rapid entrainment to spontaneous speech: A comparison of oscillator models. In N. Miyake, D. Peebles, & R.P. Cooper (Eds), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* Austin, TX: Cognitive Science Society.

Inden, B., Malisz, Z., Wagner, P., & Wachsmuth, I. (submitted). Anonymous submission.

Malisz, Z., Inden, B., Wachsmuth, I., & Wagner, P. (2012). An oscillator based modeling of German spontaneous speech rhythm. *Proceedings of Perspectives on Rhythm and Timing workshop* (pp. 38). Glasgow, UK.

Wagner, P., Inden, B., Malisz, Z., & Wachsmuth, I. (2012a). 'Ja, mhm, ich verstehe dich' - Oszillator-basiertes Timing multimodaler Feedback-Signale in spontanen Dialogen. In M. Wolff (Ed), *Elektronische Sprachsignalverarbeitung 2012 (Tagungsband ESSV) --- Studientexte zur Sprachkommunikation* (Vol. 64, pp. 179–187). Dresden: TUD Press.

Wagner, P., & Portele, T. (1999). Two dimensions of prominence. *Proceedings of the ESCA Workshop on Dialogue and Prosody.* Eindhoven, The Netherlands.

Wagner, P., Tamburini, F., & Windmann, A. (2012b). Objective, Subjective and Linguistic Roads to Perceptual Prominence. How are they compared and why? *Proceedings of Interspeech 2012* .

Wagner, P., Malisz, Z., Inden, B., & Wachsmuth, I. (in press). Interaction Phonology – A Temporal Co-ordination Component Enabling Representational Alignment within a Model of Communication. In: I. Wachsmuth, J. de Ruiter, P. Jaecks, & S. Kopp (Eds.), Alignment in Communication: Towards a New Theory of communication. Benjamins.

Wlodarczak, M., Simko, J., & Wagner, P. (2012a). Temporal entrainment in overlapped speech: Cross-linguistic study. *Proceedings of Interspeech 2012,* Portland, USA.

Wlodarczak, M., Simko, J., & Wagner, P. (2012). Syllable-boundary effect: temporal entrainment in overlapped speech. *Proceedings of Speech Prosody 2012* (pp. 611–614), Shanghai, China.

Wlodarczak, M., Simko, J., & Wagner, P. (in press). Pitch and duration as basis for entrainment of overlapped speech onsets. In: Proceedings of Interspeech 2013, Lyon, France.

Invited conference

Céline De Looze
Trinity College Dublin

Automatic analysis of speech prosody dynamics in social interactions: challenges and applications.

A major issue in the analysis of speech prosody is the definition of the temporal span of prosodic variations. They may vary over different domains, which makes their separate analysis difficult.

For instance, a major difficulty in representing intonation and segmental duration patterns is to separate global prosodic changes (determined by variations in pitch range and speech rate) from local prosodic characteristics (e.g. determined by changes in the phonological representation of intonation). This is, however, vital for the analysis of spontaneous speech, where variations in pitch range and speech rate may convey information about the speakers' discourse intentions and emotional states.

Defining the temporal span of prosodic variations at an interactional level is all the more challenging as they result from the production of two or more interactants. Conversational interaction is a dynamic and joint activity where all speakers participate in the construction of meaning and in the establishment of social relationships. It takes place according to a cooperation principle for which participants constantly adjust, accommodate or coordinate their speech with that of their conversational partner. Interpersonal accommodation, turn-taking organisation as well as backchannels account for such adjustment processes.

Determining the dynamics, coordination and roles of these coordination mechanisms is, however, difficult to achieve as their alignment may imply a temporal delay. These global and local mechanisms may be realised at different temporal spans which may overlap, be embedded or succeed each other.

Their analysis necessitates a specific approach for which their measurement is made at different anchor points and which accounts for their individual and simultaneous dynamics. This is all the more necessary in the aim of determining the role and impact each of these mechanisms have on the construction of discourse and in the expression and recognition of speakers' social states (or states in which an individual is when interacting with someone).

In this talk, methods and tools for the automatic analysis of speech prosody dynamics will be presented. Their relevance for the study of speech impairments in neurological disorders, communication skills in high-stress environments and human-robot interaction will also be discussed.

Indices prosodiques régionaux en français de Belgique. L'apport d'une catégorisation perceptive des données

Alice Bardiaux

alice.bardiaux@uclouvain.be
FNRS – Université catholique de Louvain

Abstract

This paper tackles the issue of prosodic features (duration and f0) of speakers from Brussels and Liège based on perceptive categorization of the data. The aim is not to describe what is fully and exclusively French spoken in Belgium, but to identify the prosodic features perceived as regionally marked and associated with a Belgian way of speaking.

1. Prosodie, variation régionale et perception

La présente étude porte sur les caractéristiques prosodiques du français parlé en Belgique, à Bruxelles et à Liège. À notre connaissance, la majorité des études s'intéressant à la variation prosodique régionale en français regroupent les productions étudiées en fonction de la ville d'origine du locuteur. Très souvent, les locuteurs parisiens sont considérés comme la variété de référence, parfois qualifiée de « standard », à laquelle on compare les productions des autres locuteurs étudiés, qui représentent la ou les variété(s) dite(s) régionale(s) (notamment Avanzi et al. 2012a, Schwab et al. 2012). Certaines études élargissent la variété de référence aux locuteurs du Nord de la France (Goldman & Simon 2007). D'autres études encore proposent une répartition des variétés étudiées sur un continuum de régionalité (Avanzi et al. 2012b), la variété parisienne restant la variété de référence. Dans toutes ces études, les variétés, qu'elles soient « régionales » ou « standard » sont considérées comme des ensembles de productions homogènes.

À l'inverse de ces études, nous cherchons à nous départir d'une catégorisation externe des données sur une base purement géographique, qui conduirait à postuler une manière de parler systématiquement différente entre des variétés posées a priori. En effet, l'existence de variétés régionales homogènes clairement identifiables et délimitables d'autres variétés est un mythe (Francard 2010). En outre, il n'existe pas de locuteur monostylistique (Labov 1986), le degré de standardisation d'un locuteur pouvant varier d'une production à l'autre, voire même au sein d'une même production. Dans notre étude de la prosodie du français parlé à Bruxelles et à Liège, nous privilégions donc une catégorisation perceptive des données permettant l'identification de phénomènes prosodiques considérés comme régionalement marqués et donc exploités par les auditeurs dans leurs représentations linguistiques.

Les études portant sur la prosodie du français en Belgique mentionnent généralement les traits prosodiques suivants comme caractéristiques de cette variété régionale : 1) un déplacement de l'accent final de groupe sur la syllabe pénultième (Francard 2001), 2) un allongement syllabique ou vocalique en position non-allongeable, notamment des syllabes pénultièmes de groupes (Hambye 2005, Bardiaux & Boula de Mareüil 2012, Hambye & Simon 2012) et 3) la production de contours intonatifs régionalement marqués, notamment par la réalisation d'une syllabe pénultième haute (Simon 2004)[1]. Nous concentrons donc notre attention sur l'analyse des dimensions temporelle (durée)

[1] Ces deux derniers traits pouvant participer à la perception d'un accent en position pénultième. Étant donné le syncrétisme de l'accentuation finale et de l'intonation en français (Di Cristo 1999), il est difficile de décrire l'intonation indépendamment de l'accentuation.

et mélodique (f0) non pas pour fournir une définition prosodique de ce qui constitue entièrement et exclusivement le français parlé en Belgique, mais pour décrire prosodiquement ce qui est perçu par les auditeurs comme une manière de parler régionalement marquée et associée à une manière de parler en Belgique.

2. Méthode

2.1 Participants et matériel

Le corpus d'étude utilisé est constitué d'échantillons produits par 6 locuteurs masculins, 3 Bruxellois et 3 Liégeois, répartis en 2 tranches d'âge (entre 30 et 40 ans et entre 50 et 65 ans), enregistrés dans deux tâches de lecture : lecture du texte PFC et lecture d'un dialogue conçu pour l'étude de phénomènes prosodiques. Chaque enregistrement a été transcrit manuellement dans Praat (Boersma & Weenink 2013), segmenté et aligné automatiquement à l'aide d'EasyAlign (Goldman 2011). L'alignement a été vérifié manuellement par l'auteure. Une détection automatique des proéminences a été réalisée avec ProsoProm (Simon et al. 2008)[2] (ligne d'annotation 6, figure 1).

2.2 Définition des unités prosodiques

Chaque syllabe a été annotée semi-automatiquement selon sa position syllabique (ligne d'annotation 1, figure 1) : syllabe initiale de mot plein (i), syllabe médiane de mot plein (m), syllabe finale de mot plein (f), syllabe de mot plein monosyllabique (1), syllabe de clitique (c), schwa post-tonique (@). Une tire reprenant les groupes accentuels (GA), constitués d'un mot plein et des clitiques associés (Mertens 2009 : 26), a été créée automatiquement à partir de l'annotation de la structure accentuelle des syllabes (ligne d'annotation 2, figure 1). À partir de cette tire GA et de l'annotation automatique des proéminences, une tire générant les groupes intonatifs (Mertens 2009) a été créée. Une frontière de groupe intonatif (GI) a été créée automatiquement lorsqu'une proéminence correspondait à la syllabe finale d'un GA, ou lorsqu'une proéminence était détectée à la fois sur la syllabe pénultième du GA considéré et sur la syllabe initiale du GA suivant (ligne d'annotation 3, figure 1). Ce dernier cas de figure permet de tenir compte de l'impact rétroactif d'une proéminence initiale sur la perception d'une frontière de groupe prosodique (Astésano et al. 2012).

2.3 Annotations perceptives

L'auteure a réalisé une annotation perceptive de la fonction (continuatifs vs. finaux) et de la forme (bas plat, descendant, médian plat, montant, haut plat) des contours intonatifs portés par la clausule[3] (Carton 1977 : 79) de chaque GI (ligne d'annotation 4, figure 1). Deux experts (tous deux originaires de la même région en Belgique : le Brabant Wallon[4]) ont réalisé une annotation perceptive du marquage régional de la clausule de chaque GI. Chaque syllabe de la clausule a été annotée M lorsqu'elle était perçue comme marquée régionalement et N lorsqu'elle était perçue comme non marquée régionalement (ligne d'annotation 5, figure 1). De ce codage résulte, pour chaque clausule, un code à trois lettres : NNN si aucune syllabe n'a été perçue comme régionalement marquée, NNM si seule la syllabe finale a été perçue comme marquée régionalement, NMM, si les syllabes pénultième et finale ont été perçues comme régionalement marquées, etc. L'accord inter-annotateurs est très bon puisqu'un score Kappa moyen de 0,78 est

[2] Les hésitations et les interruptions, annotées respectivement Z et !, ont été exclues des analyses (ligne d'annotation 7, figure 1).

[3] Di Cristo & Hirst (1996 : 220) soulignent l'importance des deux dernières syllabes d'unités intonatives en définissant la cadence porteuse d'une « configuration mélodique particulière », le contour intonatif.

[4] Ni Liège, ni Bruxelles ne se situent dans cette région de Belgique.

obtenu pour l'ensemble des annotations. Un code simplifié a été déduit automatiquement de cette annotation manuelle : les clausules dont au moins une syllabe a été perçue comme régionalement marquée ont été annotées M, les clausules dont aucune syllabe n'a été perçue comme régionalement marquée ont été annotées N. Après une comparaison de l'annotation perceptive des deux experts, seules les syllabes (et donc les clausules) perçues comme régionalement marquées par les deux experts ont été considérées comme régionalement marquées dans les analyses.

a	l	9~	m	d	i	z	E	l	o	t	R	@	Z	u	R		
a		l9~m		di		zE		lo		tR@			ZuR				
i		f		i		f		1		@			1				1
		if		if				1@					1				2
		if		if				1@					1				3
		$cont_h$		$cont_h$				$cont_d$					fin_b				4
		M															5
		P				P				P			P				6
																	7
	alain		me		disait				l'autre					jour			

Figure 1. Transcription, segmentation et annotations morpho-prosodiques pour la séquence « Alain me disait l'autre jour »

3. Résultats

3.1 Perception du marquage régional

Les clausules perçues comme régionalement marquées sont rares : elles représentent 7% de l'ensemble des clausules. Cette catégorisation perceptive permet d'éviter que les mesures saillantes des clausules marquées s'estompent dans la masse des mesures des clausules non marquées ; par conséquent, cette méthode permet de mettre au jour les traits prosodiques distinguant les clausules perçues comme régionalement marquées de celles qui ne le sont pas. Cette approche perceptive de catégorisation des données conduit à l'identification des traits prosodiques perçus et donc exploités par les locuteurs dans la construction de leurs représentations linguistiques.

Les syllabes le plus souvent perçues comme régionalement marquées sont les syllabes pénultièmes (χ^2 = 54,410 p < 0,001) : parmi l'ensemble des clausules perçues comme régionalement marquées, la syllabe pénultième est perçue comme marquée dans 68% des cas (figure 2).

	fréquence	pourcentage
MNN	3	3,8
NMM	8	10,2
NNM	22	28,2
NMN	45	57,8
Total	78	100

Figure 2. Fréquence et pourcentage de clausules perçues comme marquées par la syllabe antépénultième (MNN), pénultième et finale (NMM), finale (NNM) et pénultième (NMN)

Le marquage régional perçu de la clausule varie en fonction de la forme des contours intonatifs portés par la clausule (χ^2Wald (4, n = 1120) = 31,64 p < 0,001). En effet, les tests post-hoc ont montré que les contours descendants et montants présentent significativement plus de clausules perçues comme marquées (respectivement 13,8% et 12,2%) que les contours bas (4,6%), hauts (3,7%) et médians (3,1%) (figure 3).

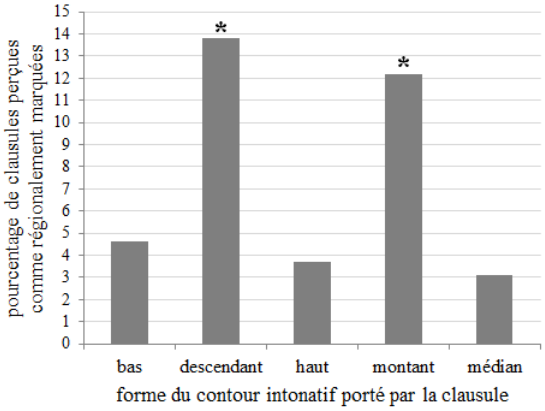


Figure 3. Pourcentage de contours marqués dans les contours b(as), d(esdendants), h(auts), m(ontants) et médians (n)

3.2 Position syllabique et marquage régional

La syllabe pénultième occupe une place prépondérante dans la perception du marquage régional. Notre attention se focalisera donc sur cette position syllabique dans l'analyse des paramètres prosodiques de durée et de f0 des clausules perçues comme marquées et non marquées régionalement afin de déterminer si la syllabe pénultième présente des mesures prosodiques significativement différentes

dans les clausules marquées et non marquées.

Durée

Les données ont été analysées au moyen d'un modèle linéaire généralisé (à mesures répétées), avec la durée syllabique comme variable dépendante et avec le marquage régional perçu et la position syllabique comme prédicteurs. Dans notre corpus, les clausules perçues comme régionalement marquées présentent effectivement des syllabes significativement plus longues d'en moyenne 23 ms[5] que les syllabes perçues comme régionalement non marquées (χ^2Wald (1, n = 3360) = 5,602 p < 0, 05).

La durée syllabique ayant tendance à augmenter au fur et à mesure qu'on approche d'une frontière droite d'un groupe prosodique (Zellner 1996), nous pouvons nous attendre à une influence de la position de la syllabe au sein de la clausule sur sa durée. Sans surprise, la durée moyenne des syllabes est en effet significativement différente selon la position syllabique (antépénultième, pénultième ou finale) de la syllabe au sein de la clausule (χ^2Wald (2, n = 3360) = 458,906 p < 0,001) : la syllabe finale est la plus longue de la clausule (plus longue de 84 ms par rapport à la syllabe pénultième, et de 107 ms par rapport à la syllabe antépénultième), suivie de la syllabe pénultième (23 ms plus longue que la syllabe antépénultième) et enfin de la syllabe antépénultième, la plus courte de la clausule.

Les syllabes pénultièmes ayant été significativement plus souvent perçues comme régionalement marquées (cf. 3.1), nous nous attendons à une interaction du marquage régional et de la position syllabique sur la durée des syllabes de la clausule. Le marquage régional montre en effet une influence différente en fonction de la position syllabique (χ^2Wald (2, n = 3360) = 16,102 p < 0,001). Selon les tests post-hoc, alors que les syllabes pénultièmes des clausules marquées sont significativement plus longues (de 29 ms en moyenne) que les syllabes pénultièmes des clausules non marquées (p < 0,001) (figure 4), les syllabes antépénultièmes et finales de clausules marquées ne présentent pas de durée significativement plus longues que leurs pendants non marqués (p > 0,05).

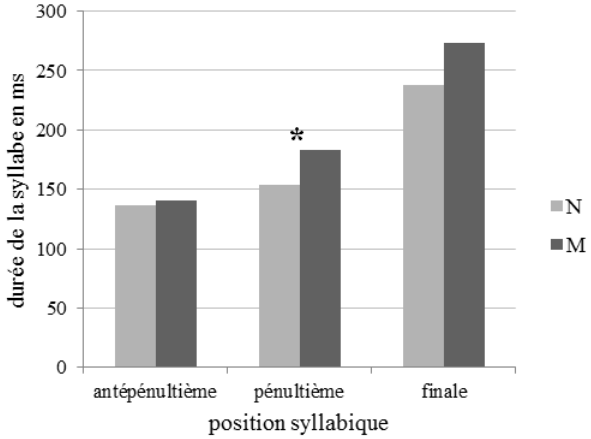


Figure 4. Durée en ms des syllabes antépénultième, pénultième et finale des clausules perçues comme régionalement marquées (M) et des clausules perçues comme régionalement non marquées (N)

Fréquence fondamentale (f0)

La f0 a également été analysée au moyen d'un modèle linéaire généralisé (à mesures répétées), intégrant cette fois trois prédicteurs : le marquage régional perçu, la position syllabique et la forme du contour porté par la clausule. Une triple interaction de ces trois facteurs sur la f0 moyenne des trois dernières syllabes du GI (χ^2Wald (4, n = 3360) = 12,784 p < 0,05) nous conduit à analyser séparément les clausules en fonction de la forme perçue de leur contour intonatif. Une analyse séparée des clausules en fonction de la forme du contour permet d'éviter que les mesures de f0 moyenne ne se compensent entre les contours hauts, montants, bas, descendants et médians. Pour chaque type de contour, l'influence du marquage régional et de la position syllabique, ainsi que leur interaction, sur la f0 moyenne des syllabes de la clausule sont analysées.

Un effet simple du marquage régional perçu est montré sur la f0 moyenne des

[5] Toutes les mesures présentées dans cet article sont les mesures réellement observées dans les données.

syllabes pour les clausules portant un contour bas (χ^2Wald (1, n = 3360) = 6,422 p < 0,05), haut (χ^2Wald (1, n = 3360) = 19,158 p < 0,001) et descendant (χ^2Wald (1, n = 3360) = 11,202 p < 0,05). La position syllabique influence également significativement la f0 moyenne des syllabes, que la clausule porte un contour bas (χ^2Wald (2, n = 3360) = 14,843 p < 0,05), haut (χ^2Wald (2, n = 3360) = 101,143 p < 0,001), descendant (χ^2Wald (2, n = 3360) = 11,277 p < 0,05), montant (χ^2Wald (2, n = 3360) = 45,920 p < 0,001) ou médian (χ^2Wald (2, n = 3360) = 6,838 p < 0,05). Une interaction existe également entre le marquage régional perçu et la position syllabique pour les syllabes des clausules portant un contour bas (χ^2Wald (2, n = 3360) = 11,093 p < 0,05) et haut (χ^2Wald (2, n = 3360) = 17, 255 p < 0,001). À l'inverse, pour les clausules portant un contour descendant, montant ou médian, le marquage régional perçu n'influence pas différemment la f0 moyenne des syllabes en fonction de leur position au sein de la clausule.

Les tests post-hoc pour les clausules portant un contour bas ont montré que les syllabes antépénultièmes et pénultièmes sont significativement plus hautes (de 10,4 Hz et 6,2 Hz respectivement) dans les clausules perçues comme régionalement marquées que dans les clausules perçues comme non marquées (figure 5). Pour les clausules portant un contour haut, ce sont également les syllabes antépénultièmes et pénultièmes qui sont significativement plus hautes (de 5,8 Hz et 4,4 Hz respectivement) dans les clausules perçues comme marquées que dans les clausules perçues comme non marquées. Dans les clausules avec un contour descendant, les trois syllabes sont significativement plus hautes (de 5,1 Hz, 7 Hz et 13,1 Hz respectivement) lorsqu'elles appartiennent à une clausule perçue comme régionalement marquée. Par contre, dans les clausules avec un contour montant ou médian, aucune des syllabes ne se distingue significativement selon que la clausule ait été perçue comme régionalement marquée ou non.

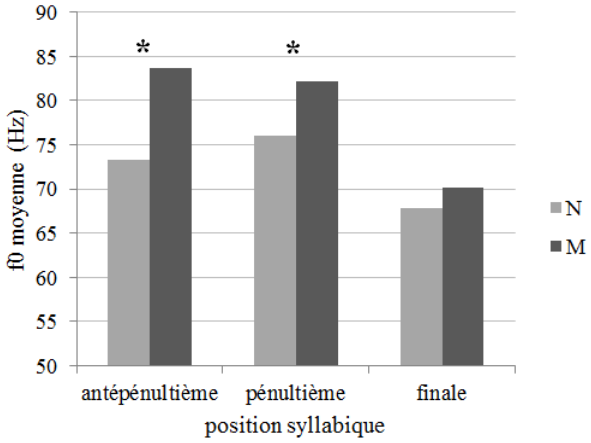


Figure 5. f0 moyenne (en Hz) des syllabes antépénultième, pénultième et finale des clausules perçues comme régionalement marquées (M) et des clausules perçues comme régionalement non marquées (N) dans les contours bas

Dans les contours bas, hauts et descendants, les syllabes antépénultième et pénultième présentent une f0 moyenne significativement différente dans les clausules perçues comme marquées que dans les syllabes perçues comme non marquées. Pour ces trois types de contours, associé au paramètre de durée, le paramètre de f0 semble donc également jouer un rôle dans la perception de segments marqués régionalement, principalement sur la syllabe pénultième. Curieusement, le marquage régional ne semble pas affecter la f0 moyenne des syllabes dans les contours montants et médians. Pourtant, les clausules portant un contour montant ont souvent été perçues comme régionalement marquées (cf. 3.1). Ce qui distingue prosodiquement les contours montants marqués des contours montants non marqués est donc à chercher ailleurs, peut-être plutôt dans le type de mouvement effectué par le contour intonatif, autrement dit, dans la manière dont les f0 moyennes des trois dernières syllabes du GI se distinguent les unes par rapport aux autres.

4. Conclusion

À l'instar des résultats obtenus par Bardiaux et al. (2012), les clausules perçues comme régionalement marquées sont rares et, parmi ces clausules marquées, la syllabe pénultième est majoritairement désignée comme responsable de ce marquage régional. La présente étude a également montré la pertinence qu'il y a à catégoriser les données en fonction de leur caractère régionalement marqué ou non, ce facteur influençant significativement à la fois la durée et la f0 des syllabes de la clausule. À l'avenir, il faudrait développer cette annotation perceptive en réunissant un plus grand nombre de juges afin d'évaluer les paramètres influençant cette perception (origine géographique des auditeurs, distance stylistique de l'auditeur avec le locuteur évalué, etc.) et, à terme, de disposer d'une annotation plus fine et nuancée, qui tiendrait donc compte de facteurs pouvant influencer l'évaluation du marquage régional.

L'analyse des paramètres prosodiques de durée et de f0 a montré que les syllabes pénultièmes de groupes intonatifs sont significativement plus longues dans les clausules perçues comme régionalement marquées que dans celles qui ne le sont pas, et ce quel que soit la forme du contour porté par les clausules. Dans les clausules portant un contour bas, haut ou descendant, cette syllabe pénultième est également significativement plus haute dans les clausules perçues comme régionalement marquées. Nos résultats vont donc dans le sens des études existantes sur la prosodie du français en Belgique en mettant en évidence l'importance de la syllabe pénultième de groupe intonatif dans la caractérisation prosodique du français perçu comme régionalement marqué et associé à une manière de parler en Belgique.

Notre méthodologie fondée sur une catégorisation perceptive des données mériterait de se voir complétée par une validation perceptive a posteriori des résultats obtenus. Cette étude gagnerait également à être étendue à un plus grand nombre de locuteurs et, surtout, à l'analyse de la parole spontanée.

Remerciements

Nous remercions sincèrement Sandra Schwab pour ses conseils dans le traitement statistique de nos données ainsi que Marie-Catherine Michaux pour son aide précieuse dans l'annotation perceptive de notre corpus.

Références

Astésano, C., R. Bertrand, R. Espesser & N. Nguyen (2012). Perception des frontières et des proéminences en français. Besacier, L., B. Lecouteux & G. Sérasset, *Actes du colloque JEP 2012*. Grenoble, pp. 353-360.

Avanzi, M., S. Schwab, P. Dubosson, J.-P. Goldman (2012a). La prosodie de quelques variétés parlées en Suisse romande. Simon, A. C. (éd.), *La variation prosodique régionale en français*. Louvain-la-Neuve, De Boeck, pp. 89-118.

Avanzi, M., N. Obin, A. Bardiaux & G. Bordal (2012b). La variation prosodique dialectale en français. Données et hypothèses. *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, vol. 1, Grenoble, pp. 457-464.

Bardiaux, A. & P. Boula de Mareüil (2012). Allongements vocaliques en français de Belgique : approche perceptive et expérimentale. *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, vol. 1, Grenoble, pp. 625-632.

Bardiaux, A., A. C. Simon & J.-P. Goldman (2012). La prosodie de quelques variétés de français en Belgique. Simon, A. C. (éd), *La variation prosodique régionale en français*. Louvain-la-Neuve, De Boeck, pp. 65-87.

Boersma, P. & D. Weenink (2013). *Praat: doing phonetics by computer* (version 5.3.42) [computer program], téléchargé le 4 mars 2013 depuis http://www.praat.org

Carton, F. (1977). Insistance dialectale : l'accent d'insistance dans les dialectes d'oïl. In : Carton, F., D. Hirst & A. Marchal (eds.) *L'accent d'insistance/Emphatic stress*. Studia Phonetica, Montréal-Paris-Bruxelles, Didier, pp. 59-92.

Di Cristo, A. (1999). Vers une modélisation de l'accentuation du français : première partie. *Journal of French Language Studies*, 9:2, pp. 143-179.

Di Cristo, A. & Hirst, D. (1996). Vers une typologie des unités intonatives. *Actes du colloque JEP 1996*. Avignon, pp. 219-222.

Francard, M. (2001). L'accent belge : mythes et réalités. Hintze, M.-A., T. Pooley & A. Judge (éds), *French Accents. Phonological and Sociolinguistic Perspectives*. London, CiLT/AFLS, pp. 251-268.

Francard, M. (2010). Variation diatopique et norme endogène. Français et langues régionales en Belgique francophone. *Langue française*, 167:3, pp. 113-126.

Goldman, J.-P. (2011). EasyAlign : an automatic phonetic alignment tool under Praat. *Actes du colloque international InterSpeech*. Septembre 2011, Florence, Italie.

Goldman, J.-P. & Simon, A. C. (2007). La variation prosodique régionale en français (Liège, Vaud, Tournai, Lyon). Description outillée. Communication aux *Journées PFC*, Paris.

Hambye, P. (2005). *La prononciation du français contemporain en Belgique. Variation, norme et identités.* Thèse de doctorat, Université catholique de Louvain.

Hambye, P. & A. C. Simon (2012). The variation of pronunciation in Belgian French: from segmental phonology to prosody. Guess, R., C. Lyche & T. Meisenburg (éds), *Phonological variation in French: illustrations from three continents*. Amsterdam, John Benjamins, pp. 129-149.

Labov, W. (1986). Language structure and social structure. Lindenberg, S., J. Coleman & S. Nowak (éds), *Approaches to Social Theory*. New York, Russell Sage Foundation, pp. 265-290.

Mertens, P. (2009). Prosodie, syntaxe et discours : autour d'une approche prédictive. Yoo, H.-Y. & É. Delais-Roussarie (éds.), *Actes du colloque IDP 2009*. Paris, 9-11 septembre 2009, pp. 19-32.

Schwab, S., M. Avanzi, J.-P. Goldman, P. Montchaud & I. Racine (2012). An acoustic study of penultimate accentuation in three varieties of French. *Proceedings of Speech Prosody 2012,* vol.1, Shanghai, pp. 266-269.

Simon, A. C., M. Avanzi & J.-P. Goldman (2008). La détection des proéminences syllabiques. Un aller-retour entre l'annotation manuelle et le traitement automatique. Durand, J., B. Habert & B. Laks (éds), *Congrès Mondial de Linguistique Française*. Paris, Institut de Linguistique Française, pp. 1685-1698.

Simon, A. C. (2004). Le domaine de la variation prosodique régionale. Aspects phonologiques et phonétiques du français parlé à Liège. Communication aux *Journées PFC*, Paris.

Zellner, B. (1996). Structures temporelles et structures prosodiques en français lu. *Revue Française de Linguistique Appliquée*, 1, Paris.

Modéliser l'interface intonosyntaxique : ratio et synchronisation entre périodes intonatives et unités illocutoires

Julie Beliao, Sylvain Kahane, Anne Lacheret

MoDyCo - UMR-7114, Université Paris-Ouest

julie@beliao.fr, sylvain@kahane.fr, anne@lacheret.com

Abstract

The role of prosody and syntax in identifying basic discourse units is a recurring issue in the studies of spoken language. In this paper, we focus on the terminal units of both syntactic and prosodic structures, studying respectively illocutionary units and intonative periods. This study focuses on their interactions and more specifically on both the synchronization and the relative number of their boundaries. Based on the analysis of a large spoken french corpus, we identify different types of such synchronizations (total, partial or absent) and relative proportions. We interpret these results from a functional point of view as an interaction between intono-syntax and the genre of discourse. Remarkably, evidence strongly suggests a much more complex intercation between syntax and prosody than expected under intuitive assumptions. Hence, we believe that the features we propose may be interesting for the study of spoken discourse.

1. Introduction

Le rôle respectif de la prosodie et de la syntaxe dans l'identification des unités de discours élémentaires constitue une question récurrente dans les travaux sur l'oral. Dans la continuité des études du GARS et des travaux effectués dans le cadre du TREEBANK RHAPSODIE, un corpus échantillonné en différents genres discursifs et annoté en prosodie et en syntaxe pour modéliser l'interface intono-syntaxique en français parlé [1], nous abordons cette question en nous focalisant à l'interface de la macrosyntaxe et de la prosodie.

Nous proposons une architecture structurée autour de ces deux niveaux de représentation, construits séparément l'un de l'autre, le premier qui s'appuie sur des contraintes distributionnelles et des tests syntaxiques pour identifier les unités qui le composent, indépendamment des informations prosodiques, le second, fondé sur des critères exclusivement perceptifs et acoustiques [2]. Cette méthode strictement modulaire ne préjuge en rien de la réalité cognitive des processus, tout au contraire : elle permet d'éviter la circularité pour mieux répondre à cette question de l'interactivité des deux composantes en situation d'interaction verbale.

Dans cette communication, nous nous centrerons uniquement sur les unités terminales des représentations syntaxiques et prosodiques manipulées dans le treebank RHAPSODIE, respectivement l'unité illocutoire (UI) et la période intonative (PI) telles qu'elles ont été définies par Lacheret et al. dans [Lacheret-Dujour et al., 2011]. Nous proposons une étude de leurs interactions dans le corpus RHAPSODIE et plus précisément de la coïncidence de leurs frontières.

Nous appelons frontières *synchronisées* les frontières qui sont à la fois des frontières de PI et d'UI. Les frontières sont *désynchronisées* quand frontières prosodiques et syntaxiques ne tombent pas en même temps, c'est-à-dire quand les frontières de PI ne sont pas des frontières d'UI et *vice versa*. Il est important de souligner que rien n'empêche qu'une PI dont la frontière s'aligne sur une frontière d'UI contienne plusieurs UI et inversement (voir la Fig-

1. http ://www.projet-rhapsodie.fr/

2. Pour une méthodologie similaire voir [Degand and Simon, 2009].

ure 1). Enfin chaque PI et UI peuvent être synchronisées *totalement* lorsque leurs frontières droites et gauches sont synchronisées ou *partiellement* lorsque c'est seulement les frontières droite ou gauche qui sont synchronisées.

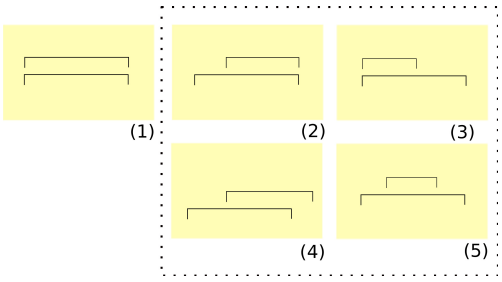


FIGURE 1. Synchronisation des frontières. (1) frontières synchronisées totalement (les deux unités s'alignent), (2) synchronisation partielle droite, (3) synchronisation partielle gauche, (4) désynchronisation chevauchante, (5) désynchronisation inclusive. En sachant que (4) et (5) sont des dérivés possibles de (2) et (3).

Notre objectif est d'abord de rendre compte statistiquement des différents types de synchronisation (totale vs partielle). Nous en proposons ensuite une lecture fonctionnelle en voyant s'il est possible de corréler les combinaisons intono-syntaxiques observées (±synchronisée) aux genres de discours à l'œuvre dans notre corpus.

2. Corpus

2.1. Corpus design

Le treebank RHAPSODIE a été créé avec comme objectif principal de proposer et de tester sur une couverture large de constructions de nouvelles méthodes d'annotation et d'analyse pour modéliser l'interface intonosyntaxique en français parlé. Un tel objectif nous a conduit à récolter à partir de données existantes (notamment les corpus présentés dans [Laks et al., 2009], [Avanzi et al., 2010] et [Branca-Rosoff et al., 2009]) et de données enregistrées pour le projet une collection d'extraits suffisamment diversifiés du point de vue de la couverture typologique (voir Table 1).

Type de parole	Monologues (M), Dialogues (D)
Situation de communication	Privée (0,1), Pro. (2)
Planification	Spontané, semi-spontané, planifié
Interactivité	Int., semi-int., non-int.
Canal	Multimédia, face-à-face
Type de séquence	Arg., descr., procédurale, ...
Tâche	cours, description, itinéraire, ...

TABLE 1. Variables situationnelles dans RHAPSODIE. Soit, 57 échantillons ; 52 hommes ; 35 femmes ; durée = 3h18 ; 34361 mots. Où : les monologues sont codés M, les dialogues D, la parole privée (0, ou 1), la parole publique (2). Ainsi, le fichier D2006 est un dialogue public, '006' indique le n° d'ordre.

2.2. Annotation prosodique

La structure prosodique associée à chaque échantillon est une structure hiérarchique composée de 5 niveaux de constituance avec de bas en haut la syllabe, le pied métrique, le groupe rythmique, le paquet intonatif et la période (voir [Lacheret-Dujour et al., 2011]). Si les trois premiers et leurs différents types sont annotés sur les bases de proéminences et des disfluences perçues, la dernière, celle qui nous intéresse ici, est segmentée semi-automatiquement sur les bases de la méthode développée par Lacheret et Victorri [Lacheret and Victorri, 2002].

En pratique, la fin d'une période est détectée si et seulement si les quatre conditions suivantes sont remplies : présence d'une pause silencieuse d'au moins 300 ms, détection d'un mouvement de F0 qui atteint une certaine amplitude, définie comme la différence de hauteur entre le dernier extremum et la moyenne de F0 sur toute la portion du signal précédant la pause, détection d'un saut, défini comme étant la différence de hauteur entre le dernier extremum de F0 précédant la pause et la première valeur de F0 suivant la pause. Il convient de souligner que la décision de reconnaissance d'une rupture périodique repose sur un principe de compensation de seuils. En d'autres termes, la détection ne dépend pas des valeurs exactes des paramètres, mais de leurs seuils respectifs d'activation et des poids associés [3] : quand

3. Activation très forte : poids '2', forte : '1', moyenne : '0', en dessous du seuil : '-1'.

un paramètre est très légèrement au-dessous du seuil choisi, une frontière de période est détectée si les autres paramètres ont des valeurs au-dessus du seuil (ex. figure 2).

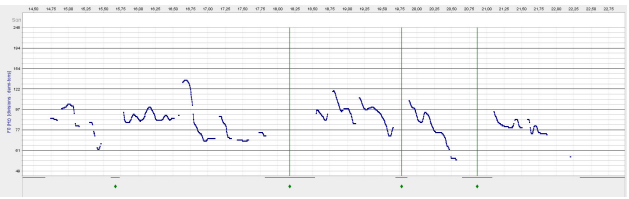


FIGURE 2. Segmentation par une barre verticale de 4 périodes de l'énoncé *je pense aux nombreuses victimes de la tempête* $[_{PI}]$ *et à toute leur famille* $[_{PI}]$ *endeuillée* $[_{PI}]$ *dont nous partageons la peine* $[_{PI}]$ [Rhap-D20004, Corpus RHAPSODIE].

2.3. Annotation syntaxique

Le système d'annotation développé pour le projet Rhapsodie comprend principalement un étiquetage morpho-syntaxique, une structure de dépendance fonctionnelle, et un découpage en unités macrosyntaxiques, que nous appelons unités illocutoires. C'est sur ce dernier que nous travaillons dans le cadre de cette étude. Le niveau macrosyntaxique[Berrendonner, 1990], [Cresti, 2000], [Benzitoun et al., 2010]définit la cohésion illocutoire à l'intérieur de l'énoncé. Cette structure est composée d'une unité centrale appelée « noyau », qui concentre la force illocutoire de l'énoncé et éventuellement d'unités satellites définies selon des critères strictement topologiques. On peut voir ci-dessous différentes configurations « satellites » possibles de l'unité illocutoire : l'exemple (1) est une construction en « pré-noyau – noyau », le (2) en « pré-noyau – noyau – post-noyau » et le (3) en « noyau – post-noyau » [4].

(1) pour eux < c'est important // [Rhap-D0006, CFPP2000][Branca-Rosoff et al., 2009]

(2) déjà < j'ai retrouvé mes origines > quand même // [Rhap-D1003]

(3) qu'est-ce que vous en pensez > de la boule magique // [Rhap-D2011]

4. La fin des pré-noyaux est indiquée par le symbole <, le début des post-noyaux par > et l'unité illocutoire se termine toujours par //.

Les modèles macrosyntaxiques déjà existants peuvent être stratificationnels ou modulaires. Le modèle définit dans le cadre du projet se différencie des modèles stratificationnels, comme par exemple celui proposé par Berrendonner [Berrendonner, 1990] qui considère les unités maximales de la microsyntaxe, comme des points d'ancrage pour la macrosyntaxe. La macrosyntaxe de Rhapsodie s'apparente plutôt aux modèles modulaires, comme celui élaboré par l'école d'Aix-en-Provence[Blanche-Benveniste, 1990], ou par Cresti [Cresti, 2000], et qui considèrent les deux types d'organisation comme orthogonaux et donc comme pouvant opérer de concert sur les mêmes séquences. En outre, le modèle de Rhapsodie considère l'organisation macrosyntaxique comme un principe de cohésion opérant de manière indépendante de la prosodie. Au final, deux critères principaux ont été retenus pour segmenter un énoncé en unités illocutoires :

– caractériser l'organisation syntaxique indépendamment de toute organisation prosodique ou du moins de tout cadre théorique prosodique ;
– proposer des critères syntaxiques [5] de segmentation explicites permettant aux annotateurs d'appliquer de manière aussi formelle que possible les choix théoriques présidant à l'annotation du corpus.

Le principal critère retenu est la non-autonomie : les segments qui ne peuvent former un énoncé autonome sont considérés comme dépendants macrosyntaxiquement.

Précisons qu'une UI peut être intégrée dans une autre UI selon deux modalités : (i) les enchâssements d'UI (le discours rapporté en (5) et les greffes [6] en (6)) notés [ ] et (ii) les insertions d'UI (les parenthèses en (7)) notées ( ).

5. La nature syntaxique de ce critère peut-être discutée notamment en regard de sa dimension pragmatique, cela-dit pour des raisons historiques nous conserverons la dénomination de « macrosyntaxe ».

6. Nous appelons greffe une UI qui vient occuper un position régie où est attendue en principe une unité lexicale. Les discours rapportés sont également des UI qui viennent occu-

(5) Marcel Achard écrivait [elle est très jolie // elle est même belle // elle est élégante //] // [Rhap-D2001, Corpus Mertens][Mertens, 1987]

(6) vous suivez la ligne du tram qui passe vers la [je crois que c'est une ancienne caserne //] // [Rhap-M0003, Corpus Avanzi][Avanzi, 2012]

(7) alors que Heinze (c'est quand même assez extraordinaire hein //) c'est le patron de la défense // [Rhap-D2003, Rhapsodie]

3. Rapports et types de synchronisations intonosyntaxiques

L'objectif de cette section est de présenter l'analyse statistique conduite sur les données en termes de rapports (équation (1)), ou relation entre les fréquences des frontières droites de PI et d'UI, et de synchronisation de ces frontières.

$$\text{Ratio}\,(\text{échantillon}) = \log\left(\frac{\text{nombre de PI}}{\text{nombre de UI}}\right). \quad (1)$$

3.1. Analyse statistique des rapports

Le corpus étudié comporte en tout 3457 UI et 2904 PI. Lorsqu'on représente pour chaque échantillon la proportion (1), on remarque que certains ressortent clairement par rapport aux autres, signifiant qu'ils présentent bien plus de PI que d'UI. Il nous est apparu de plus que ces échantillons semblaient correspondre à ceux d'un genre oratoire. Afin de vérifier statistiquement cette intuition nous avons réparti les échantillons d'après les métadonnées du corpus selon quatre genres : oratoire, procédural, argumentatif, descriptif pour lesquels il s'agit alors de vérifier si les rapports $\frac{PI}{UI}$ sont caractéristiques. À cet effet et puisque les données ne suivent pas une loi normale, c'est-à-dire qu'elles ne se répartissent pas selon une gaussienne (un test paramétrique tel qu'une ANOVA est donc proscrit), nous avons appliqué le test non-paramétrique de Kruskal-Wallis dont le résultat est affiché en figure 3. Il apparaît que l'hypothèse nulle indiquant que les rapports $\frac{PI}{UI}$ sont les mêmes dans tous les groupes peut être rejetée sans risque ($p = 5 \cdot 10^{-4}$). Nous concluons donc que ces distributions sont significativement différentes les unes des autres et par là que le calcul du rapport $\frac{PI}{UI}$ est pertinent si l'on souhaite classer et caractériser des types de discours.

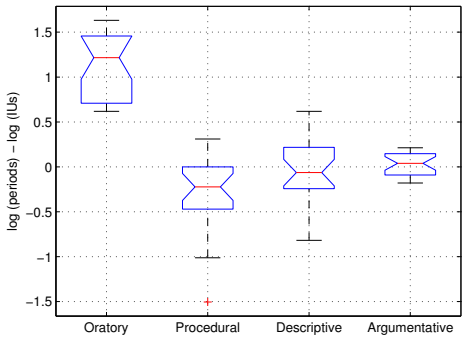


FIGURE 3. Test de Kruskal-Wallis sur les 4 groupes de sous-genres : oratoire, procédural, descriptif, argumentatif

3.2. Analyses des types de synchronisation

For de ces résultats préliminaires, nous avons vérifié si la synchronisation des frontières de PI et d'UI venait confirmer les résultats obtenus en calculant les rapports $\frac{PI}{UI}$. Pour chacun des 57 échantillons du corpus RHAPSODIE, nous avons extrait automatiquement [Belião, 2012] les frontières droites des UI et des PI ainsi que leurs positions temporelles selon différentes configurations possibles allant de le synchronisation à la désynchronisation et dont nous présentons les décomptes et pourcentages dans la table 2.

relations frontalières entre UI et PI	décompte	%
FD d'UI synchronisée avec FD de PI	1781	56
FD de PI synchronisée avec FD d'UI	1788	32
FD d'UI désynchronisées	1399	44
FD de PI désynchronisées	3799	68

TABLE 2. Décompte des relations frontalières entre PI et UI

La figure 4 présente le diagramme de dispersion des pourcentages de synchronisation des PI et des UI pour l'ensemble des échantillons.

per une position régie, mais à la différence des greffes, il s'agit d'une position où est attendue une telle construction.

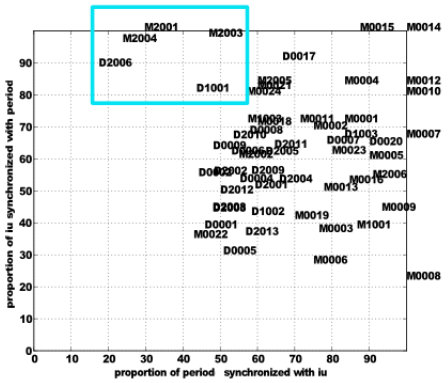


FIGURE 4. Synchronisation droite partielle des UI et des périodes pour chaque échantillons du corpus

Sur cette figure, nous observons que les échantillons appartenant au genre oratoire (encadrés) se distinguent des autres venant ainsi corroborer les résultats relatifs aux ratios présentés en figure 3. S'ils présentent une forte synchronisation des frontières d'UI avec les celles de PI (plus de 80% d'occurrences), l'inverse n'est pas vrai : dans ce genre, les périodes correspondent rarement à une UI (20 à 50%), ce qui indique qu'elles se produisent souvent au sein d'une UI et montre ainsi que la planification prosodique correspond à un processus de fragmentation syntaxique qui montre les différentes étapes de programmation de l'UI (cf (2)).

(2) lorsque vous semblez mettre en doute $\lfloor_{PI}$ notre amour des libertés $\rfloor_{PI}$ c'est un outrage $\lfloor_{PI}$ que nous n'acceptons pas $\rfloor_{UI-PI}$ nous sommes les héritiers de la tradition qui a instauré dans ce pays $\lfloor_{PI}$ la démocratie politique et sociale $\rfloor_{UI-PI}$ toujours $\lfloor_{PI}$ toujours $\rfloor_{PI}$ contre les droites coalisées $\lfloor_{PI}$ nos combats pour la conquête du droit $\rfloor_{PI}$ jalonnent l'histoire des deux derniers siècles $\rfloor_{UI-PI}$ [Rhap-D2006, Corpus RHAPSODIE]

(3) alors en partant de la place Paul Vallier pour aller à la place Notre-Dame $\rfloor_{UI-PI}$ alors j'emprunte la rue de Strasbourg $\rfloor_{UI-PI}$ je passe par la place Vaucanson $\rfloor_{UI-PI}$ je prends direction Maison du tourisme$\rfloor_{UI-PI}$ euh à la Maison du tourisme je contourne enfin je prends la rue de la République en remontant la rue de la République$\rfloor_{UI-PI}$ je tombe sur la place Sainte-Claire on va dire là où il y a la halle $\rfloor_{UI-PI}$ [Rhap-M0014, Corpus AVANZI[Avanzi, 2012]]

D'autres caractéristiques du discours pourraient être identifiées à partir de la figure 4 comme par exemple que certains échantillons situés dans le coin supérieur droit du diagramme de dispersion présentent un taux de synchronisation quasi parfait entre frontières droites d'UI et de PI (cf exemple (3)). Les locuteurs présentent un discours relativement « canonique », c'est-à-dire un discours dont la prosodie suit la segmentation syntaxique. Dans tous les cas, il s'agit de genres descriptifs (descriptions d'itinéraires) non interactifs, sans enjeux argumentatif particulier contrairement au genre oratoire mentionné plus haut.

4. Conclusion

Partant de l'étude des rapports $\frac{PI}{UI}$ (cf équation 1), les résultats observés nous ont conduit à émettre l'hypothèse qu'il s'agissait peut-être là d'un critère discriminant pour classer et caractériser les genres de discours. Pour illustrer cette hypothèse, nous nous sommes focalisés sur la question de la synchronisation des frontières droites des unités syntaxiques et prosodiques étudiées. Les points qui retiennent particulièrement notre attention sont les suivants : la différence entre la fréquence d'usage observée et la fréquence intuitive attendue qui s'illustre par une production massive de PI par rapport au nombre d'UI (3180 UI pour 5587 PI, table 2) et la relation probable entre cette répartition et les genres de discours en jeu. Si ce premier constat reste à renforcer sur du matériel plus exhaustif, et peut-être avec des descripteurs plus approfondis, il apporte une contribution significative aux études sur les genres oraux, qu'elles s'inscrivent dans le champ de la typologie textuelle ou dans celui de la phonostylistique. Il nous ouvre notamment une piste prometteuse et encore peu exploitée à l'interface de la prosodie, de la syntaxe et de la sémantique : ces traces intonosyntaxiques que les genres laissent dans le message pourraient être des marqueurs stables de la façon dont les relations de discours (contraste, concession, élaboration résumé, etc) se tissent dans les textes.

5. References

[Avanzi, 2012] Avanzi, M. (2012). *L'interface prosodie/syntaxe en francais. Dislocations, incises et*

asyndètes. Bruxelles, Peter Lang. PhD thesis, Université de Neuchatel.

[Avanzi et al., 2010] Avanzi, M., Simon, A.-C., Goldman, J.-P., and Auchlin., A. (2010). Un corpus de français parlé annoté pour l'étude des proéminences, c-prom. *Actes des 23èmes journées d'étude sur la parole, Mons, Belgique*.

[Belião, 2012] Belião, J. (2012). Formalisation, implémentation et exploitation d'une hiérarchie objet intono-syntaxique : étude sur un treebank de francais oral spontané. Master's thesis, Université Sorbonne Nouvelle.

[Benzitoun et al., 2010] Benzitoun, C., Dister, A., Gerdes, K., Kahane, S., Pietrandrea, P., and Sabio, F. (2010). Tu veux couper là faut dire pourquoi. propositions pour une segmentation syntaxique du franąis parlé. *Actes du Congrès Mondial de Linguistique franąise, La Nouvelle Orléans.*

[Berrendonner, 1990] Berrendonner, A. (1990). Pour une macro-syntaxe. *Données orales et théories linguistiques*, 21 :25–31.

[Blanche-Benveniste, 1990] Blanche-Benveniste, C. (1990). Un modèle d'analyse syntaxique 'en grilles' pour les productions orales. *Anuario de Psicologia Liliane Tolchinsky (coord.) Barcelona*, vol. 47 :11–28.

[Branca-Rosoff et al., 2009] Branca-Rosoff, S., Fleury, S., Lefeuvre, F., and Pires, M. (2009). Discours sur la ville. Corpus de Français Parlé Parisien des années 2000.

[Cresti, 2000] Cresti, E. (2000). Corpus di italiano parlato. *Florence, Accademia della Crusca.*

[Degand and Simon, 2009] Degand, L. and Simon, A.-C. (2009). *Where prosody meets pragmatics*, chapter Mapping prosody and syntax as discourse strategies : How basic discourse units vary across genres. Emerald Group.

[Lacheret and Victorri, 2002] Lacheret, A. and Victorri, B. (2002). La période intonative comme unité d'analyse pour l'étude du francais parlé : modélisation prosodique et enjeux linguistiques. *Verbum*, pages 55–72.

[Lacheret-Dujour et al., 2011] Lacheret-Dujour, A., Kahane, S., Pietrandrea, P., Avanzi, M., and Victorri, B. (2011). Oui mais elle est où la coupure, là ? Quand syntaxe et prosodie s'entraident ou se complètent. *Langue française, Paris-Larousse*, 170 :61–80.

[Laks et al., 2009] Laks, B., Durand, J., and Lyche, C. (2009). Le projet PFC (Phonologie du Français Contemporain) : une source de données primaires structurées. In *Phonologie,variation et accents du français Hermès*, pages 19–6. Hermès.

[Mertens, 1987] Mertens, P. (1987). *L' intonation du français : de la description linguistique à la reconnaissance automatique*. Katolieke universiteit te Leuven.

Le phrasé prosodique dans les variétés africaines de français

Guri Bordal[1] & Gélase Nimbona[2]

Guri.bordal@gmail.com, gelase.nimbona@uclouvain.be
[1] MultiLing (CoE), Université d'Oslo
[2] Institut Language et Communication / Université catholique de Louvain

Abstract

The aim of this study is to get a deeper understanding of the factors that determine prosodic phrasing in African varieties of French, with particular focus on the influence of African L1 in African varieties. Former studies indicate that the main difference between African varieties where French is in contact with African languages and European French is the presence of a word prosodic system in the former (Avanzi et al. 2011). Through a comparison of the distribution of syllabic prominences in four African varieties (CAR, Mali, Senegal and Burundi) and two European varieties (Paris and Brussels), we seek to find out if word prosody is a general tendency of African French. Our data show the Burundi speakers produce longer prosodic phrases than the other, which we argue is due to the important role of the Obligatory Contour Principle in Kirundi, the speakers' L1.

1. Introduction

Les études conduites jusqu'à présent sur les variétés africaines du français révèlent une différence prosodique fondamentale entre la structure de celles-ci et la structure que prédisent les modèles phonologiques de la prosodie du français, en particulier (Jun & Fougeron 2002). En fait, ces modèles, basés sur des données parisiennes, ne prédisent pas de marquage prosodique obligatoire de chaque unité lexicale (Jun & Fougeron 2002). Le français se distinguerait ainsi de la majorité des langues au monde par son système prosodique uniquement post-lexical, où le domaine de l'accent primaire est le syntagme accentuel (SA). Cette prédiction se confirme lorsqu'elle est confrontée à des données parisiennes (lectures) (Avanzi et al. 2011), mais ne semble pas valable pour des données sénégalaises, maliennes et centrafricaines où les locuteurs tendent à marquer prosodiquement chaque unité lexicale, voire certains mots fonctionnels (Avanzi et al. 2011, Bordal 2012, Bordal & Lyche 2012). Les spécificités prosodiques des variétés africaines ont été attribuées à l'influence des langues africaines (Bordal 2012), qui sont toutes des langues à prosodie lexicale: le bambara (Mali) et le sango (Centrafrique) ont des tons lexicaux tandis que le wolof (Sénégal) a un accent de mot.

Dans cette étude, nous examinerons si la tendance à marquer prosodiquement chaque mot peut être généralisée à d'autres variétés africaines, à la lumière d'une variété, jusqu'à présent non étudiée : le français burundais (désormais FBI). Nous comparons également les données africaines à une autre variété européenne, à savoir le français parlé en Belgique (désormais FBE). Cette comparaison permettrait de cerner davantage si certains traits prosodiques seraient typiquement «africains» (à la différence des traits plutôt « européens »). Les études existantes sur le FBE indiquent en effet que celui-ci possède le même type de système prosodique que le français parisien dans la mesure où la parole se découpe minimalement dans les SA (Hambye & Simon 2009, Bardiaux et al.

2012) [(1)].

A travers une comparaison de la réalisation de proéminences dans un corpus de données similaires de nos quatre variétés africaines (Sénégal, Mali, Centrafrique, Burundi) et deux variétés européennes (Paris et Bruxelles), nous verrons que la généralisation tient partiellement : le FBE se comporte selon les prédictions du modèle (Jun & Fougeron 2002) tandis que les données du FBI indiquent que cette variété a un système «intermédiaire» entre celui des varétés européennes et africaines en matière de marquage prosodique des unités.

2. Méthode

Nos analyses se basent sur la lecture du même texte, à savoir celui du projet PFC (Durand & Lyche 2002). Nous comparons la lecture de ce texte de quatre locuteurs de nos six points d'enquête. Les locuteurs ont été sélectionnés selon leur profil sociolinguistique (âge, niveau de formation, la L1 des locuteurs et sexe) [(2)].

Les enregistrements sont transcrits orthographiquement sous Praat [3] et segmentés et alignés en mots graphiques, syllabes et phonèmes à l'aide du script EasyAlign (Goldman 2011). Les analyses se basent sur la détection perceptive de proéminences syllabiques, la proéminence étant ici définie comme « une syllabe qui se détache acoustiquement et/ou perceptivement de son entourage ». Deux experts en prosodie ont écouté des extraits d'enregistrements (pas plus de 6 secondes). Ils notaient un « P » sous les syllabes qu'ils ont perçues comme très proéminentes ou un « p » sous les syllabes perçues comme proéminentes, mais moins saillantes que celles annotées « P ». Les annotations des deux experts ont été comparées pour l'élaboration d'une annotation de référence. Dans les cas de désaccord, un troisième expert est intervenu pour déterminer la nature proéminente/non proéminente de la syllabe.

Les syllabes proéminentes ont par la suite été interprétées comme frontières de groupes prosodiques [(3)]. Nous avons limité notre analyse (i) à la comparaison du nombre total de proéminences par point d'enquête et (ii) au test de deux paramètres relatifs à la formation du syntagme accentuel: (1) le respect de la contrainte dite *CLASH (Post 2000) qui interdit l'adjacence de deux accents primaires et (2) le respect de la contrainte ALIGN-Xhead [17] qui assure que les bords droits des groupes accentuels s'alignent sur des frontières droites de constituants syntaxiques X´ (Selkirk 1984).

3. Résultats

3.1. Proéminences

Le tableau 1 indique que la proportion de syllabes proéminentes (finales comme internes) est moins élevée dans le corpus de FBI que dans les autres corpus africains, alors que le FBE se comporte comme le

[1] Hambye & Simon (2009) ont en effet montré que, s'il est indéniable que certains traits rendent les pratiques linguistiques des belges francophones particulières, il n'est pas pour autant simple de modéliser un système phonologique capable de rendre compte de ce qui distingue l'accent des francophones de Belgique et celui de leurs homologues français, suisses, québécois, etc. (2009 :96).

[2] Il est en effet couramment évoqué une hypothèse selon laquelle les femmes joueraient un rôle dans la standardisation, mais la pertinence de ce paramètre ne rencontre pas l'unanimité des auteurs en contexte africain (Dister et al. 2008). En l'absence d'hypothèses spécifiques par rapport au sexe des locuteurs, la sélection de nos données ne tient pas compte de cette variable.

[3] Étant donné que (i) les proéminences qui ne sont pas en position non finale peuvent frapper n'importe quelle autre syllabe que finale du groupe (Pasdeloup, 1990 ; Avanzi, Lacheret-Dujour, Obin & Victorri (2011) en fonction des contraintes diverses, rendant leur prédiction difficile ; nous nous sommes uniquement intéressés aux frontières droites des groupes (c'est-à-dire à la distribution des accents finaux).

français parisien.

Terrain	Syllabes[4]	Proéminences	%
RCA	2485	1019	41,01%
Mali	2496	1020	40,87%
Sénégal	2468	932	37,76%
Burundi	2546	853	**33,50%**
Paris	2476	771	31,14%
Belgique	2470	741	30%

Tableau 1 : Nombre total de proéminences en ordre décroissant

La proportion des syllabes proéminentes par rapport au nombre total des syllabes produites dans chaque corpus décroît des variétés de contact (africaines) aux variétés européennes. On peut remarquer que les variétés centrafricaine et malienne d'un côté, et les variétés européennes occupent les extrémités d'un continuum au milieu duquel se situe la variété burundaise.

*3.2. Violation de la contrainte *CLASH*

Le taux de violation de la contrainte *CLASH sur la distribution des proéminences montre également que le FBI se situe entre les variétés européennes et les autres variétés africaines (cf. tableau 2).

Terrain	Occurrences[5]	Violations	%
RCA	44	39	88,64%
Mali	44	33	75,00%
Sénégal	44	25	56,82%
Burundi	44	17	**38,63%**
Paris	44	8	18,18%
Belgique	44	9	20,45%

*Tableau 2 : Violation de la contrainte *CLASH*

À titre d'exemple, ¾ des locuteurs FBI réalisent le syntagme « Marc Blanc» en un seul SA alors que dans les autres variétés africaines, ce même syntagme est systématiquement réalisé en deux.

3.3. Violation de la contrainte ALIGN-XP

Le taux de violation de la contrainte ALIGN-XP est aussi plus bas dans le corpus du FBI que celui observé dans les corpus des autres variétés africaines.

Terrain	Occurrences[6]	Violations	%
Mali	72	67	93,06
Sénégal	72	50	69,44
RCA	72	41	56,95
Paris	72	31	**43,06**
Burundi	72	34	47,22
Belgique	72	18	25

Tableau 3 : Violation de la contrainte ALIGN-XP

Exemple : le syntagme « un gros détachement » est réalisé en une seule unité prosodique par la moitié des locuteurs du corpus FBI alors que chez tous les locuteurs des autres variétés africaines étudiées, ce syntagmes est réalisé en deux SA.

4. Discussion

Comment peut-on expliquer que le FBI serait différent des autres variétés africaines alors que le kirundi tout comme le wolof, le bambara et le sango se situe dans la typologie des langues à prosodie lexicale ?

Les analyses que nous avons effectuées avèrent une tendance à ce que les locuteurs burundais, à la différence des autres locuteurs africains, tendent à réaliser moins de proéminences et font ainsi preuve d'un système proche d'un système «post-lexical» comme celui décrit dans Jun & Fougeron 2002.

[4] Le nombre de syllabes réalisées dans la lecture diffère selon les locuteurs en raison de la variation de plusieurs phénomènes phonotactiques, notamment la chute de schwa.

[5] Avec quatre locuteurs par enquête, les 11 contextes donnent 44 occasions, pour chaque terrain, d'appliquer la contrainte.

[6] Avec quatre locuteurs par enquête, les 18 contextes donnent 72 occasions pour chacun des terrains.

Si l'on s'en tient toujours au fait que la L1 des locuteurs influence leur français L2 (Bordal 2012, Bordal & Lyche 2012, Bordal et al. 2012), y aurait-il lieu d'expliquer cette divergence par les spécificités du système prosodique du kirundi par rapport à ceux des autres langues africaines parlées par les locuteurs de notre corpus ?

Le kirundi, comme le bambara et le sango, est une langue à tons (donc à prosodie lexicale). Cependant, contrairement au sango et au bambara dont les systèmes prosodiques présentent une certaine compacité tonale, le kirundi possède un système à tons retreints ; il comporte le seul niveau de hauteur – le niveau haut (H) – distinctif dans les représentations phonologiques (/H/ vs. /Ø/) (Rialland 1998). Outre cette restriction tonale dans les représentations de base, la distribution tonale en kirundi est sous-tendue entre autres par la règle dite de Meeussen ('Meeussen's rule'[7]) qui interdit la succession de deux tons H dans la même séquence (Rialland 1998), ce qui n'est pas le cas par exemple en sango (Diki-Kidiri 1977). Peut-on établir des liens entre cette règle et la contrainte *CLASH en français pour expliquer le taux réduit de violation de celle-ci en FBI par rapport aux autres variétés africaines? Si oui, expliquerait-elle le fait que le FBI tend à privilégier la segmentation du flux de parole en SA plutôt qu'en mots prosodiques, comme en français centrafricain (Bordal 2012), en français malien (Bordal & Lyche 2012) ou en français sénégalais (Bordal et al. 2012) ? Notre échantillon semble trop restreint pour confirmer nos résultats. Cependant, il importe de remarquer que la différence entre ces variétés de contact font écho aux prédictions de l'Hypothèse de Marquage Différentiel (Eckman 2004) dans le domaine d'acquisition ; à savoir que les apprenants de L1 différentes ne rencontrent pas les mêmes difficultés dans l'acquisition de la structure de la [même] L2. Le degré de variation dans la maîtrise des apprenants de certains traits de L2 dépend largement des spécificités de leur L1.

[7] La règle dite de Meeussen nous semble être le corollaire du Principe de Contour Obligatoire « Obligatory Contour Principle (OCP) dans la théorie autosegmentale qui interdit l'adjacence de deux autosegments identiques (Goldsmith, 1976) ; mais comme les tons B(as) et M(oyen) ne sont que des détails phonétiques en kirundi dont la valeur n'est évaluée que par rapport à la présence du ton H(aut), les deux règles sont applicables.

5. Conclusion

Dans cette étude, nous avons montré que les réalisations des locuteurs de notre corpus burundais se distinguent de celles autres locuteurs africains que nous avons étudiées. Pour tous les phénomènes étudiés – (1) la proportion des syllabes proéminentes ainsi que (2) le respect des contraintes *CLASH et ALIGN-Xhead – les résultats montrent que les réalisations des locuteurs burundais se distinguent des autres locuteurs africains en matière du phrasé prosodique. Notre hypothèse est que cette différence peut être attribuée à la spécificité du système prosodique du kirundi (L1 des locuteurs burundais) par rapport à ceux des L1 des locuteurs des autres variétés. Une analyse de plus de locuteurs ainsi que de la parole spontanée serait de rigueur pour mieux explorer cette problématique. Il serait également intéressant pour un travail ultérieur d'étudier de plus près les aspects sociolinguistiques, c'est-à-dire de comparer la situation de contact linguistique du français différente dans ces pays, afin de discuter si le mode d'acquisition et les contextes d'utilisation du français peuvent y jouer un rôle.

Références

Avanzi, M, Bordal, G. & Obin, N. (2011). Variations in the Realization of the French Accentual Phrase. *Proceedings of ICPhS*, Hong Kong, China.

Avanzi, M., A. Lacheret-Dujour, N. Obin & B. Victorri (2011). Vers une modélisation continue de la prosodie: le cas des proéminences

syllabiques. *Journal of French Language Studies* 21:1, pp. 53-71.

Bardiaux, A., A.-C. Simon & J.-P. Goldman, (2012). La prosodie de quelques variétés de français parlées en Belgique. Simon, A.C. (éd.), *La variation prosodique régionale en français*, De Boeck , Bruxelles, pp. 65-88.

Boersma, P. & Weenink, D. (2012). Praat, v.5.3. http://www.fon.hum.uva.nl/praat/.

Bordal, G. (2012). *Prosodie et contact de langues: le cas du système tonal du français centrafricain*. Thèse de doctorat, Université d'Oslo/Université de Paris Nanterre.

Bordal, G. & C. Lyche (2012). Regard sur la prosodie du français d'Afrique à la lumière de la L1 des locuteurs. Simon, A.C. (éd.). (2012). *La variation prosodique régionale en français*, De Boeck, Bruxelles, pp. 179-198.

Bordal, G, M. Avanzi, N. Obin, & A. Bardiaux (2012). Variations in the realization of French accentual Phrase in the light of language contact. *Proceedings of Speech Prosody*, Shanghai, China.

Diki-Kidiri, M. (1977). *Le sango s'écrit aussi: esquisse linguistique du sango, langue nationale de l'empire centrafricain*. Société d'études linguistiques et anthropologiques de France, Paris.

Dister, A., F. Gadet, R. Ludwig, C. Lyche, L. Mondada, S. Pfänder, A. C. Simon & I. Skattum (2008). Deux nouveaux corpus internationaux de français : CIEL-F (Corpus International et écologique de la langue française et CFA (Français contemporain en Afrique et dans l'Océan indien). *Revue de Linguistique Romane* 72, pp. 295-314.

Durand, J. & C. Lyche (2002). La phonologie du français contemporain: usages, variétés et structure. Delais-Roussarie, E. & J. Durand, (éds.), *Corpus et Variation en phonologie du français.* Presses universitaires de Mirail, Toulouse, pp. 213-276.

Eckman, F. (2004). From phonemic differences to constraint rankings, *SSLA* 26, Cambridge University Press, Cambridge, pp. 513–549.

Goldman, J.-P. (2011). EasyAlign: an Automatic Phonetic Alignment Tool under Praat. *Proceedings of Interspeech* XI, pp. 3233–3236, Université de Genève, Genève.

Goldsmith, J.A. (1976). *Autosegmental Phonology*. Ph.D. Thesis, MIT.

Hambye, P. & A.C. Simon, (2009). La prononciation du français en Belgique. J. Durand, B. Laks & C. Lyche (éds.) *Phonologie, variation et accents en français*, Hermès, Paris, pp. 95-130.

Jun, S.-A., & C. Fugeron (2002). Realizations of Accentual phrase in French intonation. *Probus* 14, pp. 147-172.

Pasdeloup, V. (1990). *Modèle de règles rythmiques du français appliqué à la synthèse de la parole.* Thèse de doctorat, Université de Provence Aix-Marseille.

Post, B. (2000). *Tonal and Phrasal Strucures in French*. Holland Academic Graphics,The Hague.

Rialland, A. (1998). Systèmes prosodiques africains : une source d'inspiration majeure pour les théories phonologiques multilinéaires. S. Platiel et R. Kaboré (éds.), *Langues africaines subsahariennes* Numéros spéciaux de '*Faits de langues*, pp. 407-428.

Selkirk, E. (1984). *Phonology and Syntax: the relation between Sound and Structure*. Cambridge (MIT Press), Cambridge.

Prosodic features of simultaneous interpreting

George Christodoulides

george@mycontent.gr
Université de Louvain, Louvain-la-Neuve, Belgium

Abstract

We study the prosodic features of simultaneous conference interpreting in an attempt to describe its particular speaking style. We focus on the temporal organisation of the interpreters' speech (pauses, speech rate), as well as global prosodic properties (f_0 range, melodic agitation). We also study similarity and convergence phenomena between the speaker and the interpreter, on prosodic features, and their dynamic evolution over time. The findings indicate that interpreters make longer silent pauses, less frequently than speakers and their speech rate is more variable. In most cases, interpreters had a narrower pitch range than speakers and do not mirror the pitch of their speakers.

1. Introduction

Simultaneous interpreting (SI) has been described as a taxing cognitive task, during which the interpreter is working at the limits of their processing capacity (Pöchhacker 2004). Speaking style (*phonostyle*) is determined both by the situational context and by individual characteristics (Llisteri 1992; Eskenazi 1993; Léon 1993; Simon et al. 2010). The interpreter may choose to alter some local prosodic characteristics of their speech to mimic choices made by the speaker (Couper-Kuhlen & Selting 1996) if they deem it appropriate. The interplay of all these factors will define the speech style of the interpreter. Some interpreters may adopt a uniform, personal style regardless of the speaker, while others will be more influenced by and converge with the speaker.

This study focuses on conference interpreting of speeches delivered in public, in a political context. Our hypothesis is that the time and cognitive constraints of SI create a specific speaking style of interpreting. We also investigate whether and to what extent the prosodic characteristics of an interpreter's speech are influenced by those of the speaker and their evolution over time.

2. Corpus

2.1 Design and structure

A parallel bilingual spoken corpus was built for this study, consisting of speeches given in English at the European Parliament, and their interpreted versions in French. We have chosen to focus on one situational context, i.e. argumentative political discourse in the EU institutions, to avoid possible variations due to different contexts.

One sub-corpus consists of interventions at committee meetings, and another contains short speeches in the plenary. Since the rules of procedure place more stringent time constraints to speakers in the plenary, their succession is faster and the debates tend to be livelier. An additional corpus of two press conferences was used for comparison.

In order to study individual variation across speakers and interpreters, the corpus is organised in cross-tabulation groups: each speaker in a given group has been interpreted by every interpreter in the group, and vice versa, as summarised in Table 1.

Committee meetings			**Plenary sessions**		
Spk	**Int**	**Dur (s)**	**Spk**	**Int**	**Dur (s)**
S1en	I1fr	2 x 123	S4en	I3fr	2 x 330
S1en	I2fr	2 x 129	S4en	I4fr	2 x 228
S2en	I1fr	2 x 178	S5en	I3fr	2 x 170
S2en	I2fr	2 x 142	S5en	I4fr	2 x 126
S2en	I1fr	2 x 155	**Press conferences**		
S3en	I2fr	2 x 225	S6en	I5fr	2 x 205
S3en	I1fr	2 x 186	S6en	I6fr	2 x 273

Table 1: Corpus design and sample durations

Each of the 13 corpus samples is a time-synchronised recording of the original (EN) and the interpretation (FR). The total corpus duration is 82.5 minutes (2474 s or 41.3 min per language), with 6 different speakers (5 male, 1 female) and 6 different interpreters (2 male, 4 female).

2.2. Corpus annotation

All corpus samples have been transcribed, phonetised and aligned to the phone level using *SPPAS* (Bigi 2012) and *EasyAlign* (Goldman 2008). We performed syllabification with *SPPAS* for French, and our own reimplementation of the *P2TK syllabifier* for English. The phonetic alignment was manually corrected. The data is stored in *Praat* (Boersma & Weenink 2009) textgrids.

Prosodic information was extracted by applying f_0 stylisation using *ProsoGram* (Mertens 2004), *ProsoReport* (Goldman et al. 2011), and automatic prominent syllable detection (Goldman et al. 2012).

Bi-text alignment based on translational equivalence was performed on pause-separated units. This results in parallel macro-units of English speech and the corresponding French interpretation.

We have developed software to manage such a parallel spoken corpus, perform automated analyses and visualise the results, partly based on the open source project *Sonic Visualiser* (Cannam et al. 2010).

3. Temporal features of SI

The temporal organisation of the interpreter's speech depends upon the speaker and the reformulation process. The interpreter has to obtain enough source language input before starting to produce a coherent message in the target language. Gile (2009: 200) argues that interpreters use several strategies to cope with the cognitive load of this process, including stalling ('delaying the response'), varying the ear-voice span and their speech rate, and anticipating. These strategies are reflected in the observed temporal features of interpreters' speech.

Overall, interpreters make longer and less silent pauses than speakers. The distribution of silent pause durations is presented in Figure 1.

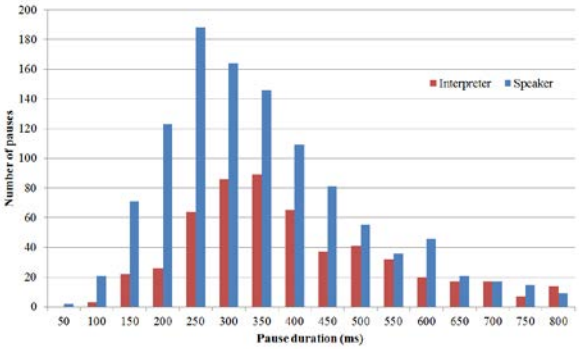


Figure 1: Silent pause duration distribution (blue: speaker / red: interpreter). All situations.

We also observe differences based on the situation: in the quicker debates at the plenary, the number of pauses and their average duration is smaller for speakers and interpreters alike, but the different patterns persist. At the press conferences, the speaker made more pauses than the interpreter, with similar pause durations.

	Mean (ms)	Medn (ms)	Std Dev	Avg num pauses /min
Speaker	**322**	**304**	**157**	**25.7**
Committee	328	311	152	20.3
Plenary	295	249	157	37.6
Press conf.	510	479	265	20.8
Interpreter	**594**	**379**	**640**	**13.8**
Committee	707	445	776	12.4
Plenary	412	329	319	18.2
Press conf.	827	508	962	10.0

Table 2: Silent pause length and frequency

Following the methodology in Goldman et al. (2010), the silent pause length distributions were modelled using a Gaussian mixture model in log time. Both were found to be bi-modal using a BIC criterion. The models are as follows (all times in ms):
For speakers (λ_1=8%, μ_1=105, σ_1=1.39 and λ_2=92%, μ_2=314, σ_2=1.49)

For interpreters (λ_1=45%, μ_1=329, σ_1=1.31 and λ_2=55%, μ_2=569, σ_2=2.26).

We also observed that in most (8 in 11) samples, and overall, the variance of speech rate was greater for speakers than for interpreters, which suggests that interpreters accelerate and decelerate more than speakers.

4. Global prosodic profile

	Committee				Plenary	
	S1	**S2**	**S3**		**S4**	**S5**
Agitation (semitones / second)						
Spk	7.60	8.35	6.60	**Spk**	4.80	7.20
I1	4.70	5.40	4.50	**I3**	4.40	4.90
Spk	7.10	8.80	4.40	**Spk**	4.90	7.30
I2	6.20	6.60	6.30	**I4**	5.80	5.60
F0 range (5%-95% interquartile range, semitones)						
Spk	7.80	12.30	12.10	**Spk**	11.10	12.90
I1	7.30	7.30	7.00	**I3**	8.20	9.40
Spk	7.40	15.00	7.90	**Spk**	11.20	13.70
I2	18.70	9.40	20.40	**I4**	5.70	6.20
Articulation ratio (%)						
Spk	84.7	89.3	81.9	**Spk**	79.9	79.4
I1	90.8	88.3	83.3	**I3**	82.8	81.8
Spk	84.1	89.2	81.7	**Spk**	79.6	79.0
I2	78.0	83.2	68.0	**I4**	90.5	92.6
Articulation rate (syllables / s, excluding pauses)						
Spk	6.10	4.95	4.90	**Spk**	5.10	5.40
I1	4.20	3.95	3.70	**I3**	4.80	4.80
Spk	5.70	5.10	4.60	**Spk**	5.70	5.00
I2	3.60	3.60	3.70	**I4**	4.90	4.40
Speech rate (syllables / s, including pauses)						
Spk	5.12	4.40	4.01	**Spk**	4.09	4.28
I1	3.83	3.48	3.06	**I3**	3.96	3.91
Spk	4.80	4.57	3.76	**Spk**	4.52	3.95
I2	2.78	2.98	2.51	**I4**	4.48	4.03

Table 3: Global prosodic measures

Interpreters have a lower average speech rate (including pauses) and articulation rate (excluding pauses) than speakers at the plenary and at committees, while these two rates are roughly equal at press conference. Interpreters' speech rate and articulation rate have greater variance than the corresponding rates for speakers. Most interpreters displayed a narrower f_0 range than speakers, and a lower melodic agitation (as defined in Goldman et al. 2007: 225), which suggests that they do not emphasise their speech as much as speakers do.

Finally, we can observe that within these general patterns, individual interpreters have their own speaking style. For example, I2 systematically uses a broad pitch range, regardless of the speaker. These observations lead us to study the dynamic evolution prosodic features, comparing the speaker and the interpreter over time.

5. Similarity and convergence

5.1 Methodology

We have studied the evolution of three prosodic features (speech rate, mean pitch and pitch range) over time based on the Time-Aligned Moving Average (TAMA) method, described by Kousidis et al. (2009). This method has been used to identify similarity and convergence in spontaneous dialogues. De Looze & Rauzy (2011) propose measures for synchrony (two speakers exhibiting similar speech patterns) and convergence (moving towards similar prosodic features over the course of time).

In simultaneous interpreting the two time series are naturally aligned, and since there is no interaction, only the interpreter can converge towards the prosodic features of the speaker.

For each prosodic feature under study, its value was extracted every 1 s, creating two time series that were normalised using the z-transformation. The moving average was calculated with a window size of 10 s. The synchrony (S) and convergence (C) strengths are calculated using Pearson's rho:

$$S = \rho_{\text{Pearson}}(x_1, x_2), \quad C = -\rho_{\text{Pearson}}(|x_1 - x_2|, t)$$

where x_1 and x_2 are the normalised TAMA values. Synchrony and convergence are independent, giving seven possible states: 3 states of similarity, 3 states of anti-similarity and 1 state of no similarity; see De Looze & Rauzy (2011). The dynamic nature of the phenomenon is captured by observing the plots of the two series and the two indices (S, C) over time.

5.2 Regions of synchrony

The interpreter and the speaker may be in synchrony on one prosodic parameter (e.g. speech rate) and at the same time in anti-synchrony on another one (e.g. pitch range). Table 4 summarises the percentage of time each state was observed.

	Speech rate			Pitch range		
	ANT	**NIL**	**SYN**	**ANT**	**NIL**	**SYN**
S1-I1	23	46	31	19	36	45
S2-I1	25	49	26	17	46	37
S3-I1	30	57	13	24	54	22
S1-I2	28	31	41	28	40	32
S2-I2	11	25	64	33	58	9
S3-I2	29	39	31	37	27	36
S4-I3	37	30	34	15	44	41
S5-I3	15	65	20	16	44	40
S4-I4	10	55	36	15	49	36
S5-I4	46	27	27	10	46	44

Table 4: Percentage of time (%) the speaker and the interpreter are in synchrony (SYN) or anti-synchrony (ANT) phases ($p<0.05$)

5.3. Speech rate

Speech rate presents the most interesting variation, as it is linked with the temporal constraints of simultaneous interpreting. We find alternating regions of synchrony (i.e. the interpreter is following the changes in speech rate of the speaker) and regions of anti-synchrony.

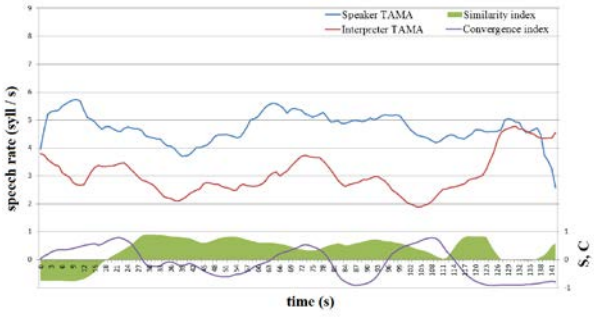


Figure 2: Speech rate TAMA plot with the interpreting following the speaker (S2/I2).

Two distinct styles emerge: the interpreter may be following quite faithfully the accelerations and decelerations of the speaker (e.g. Figure 2) or may be keeping their own pace (e.g. Figure 3). The second style was observed in most of the samples.

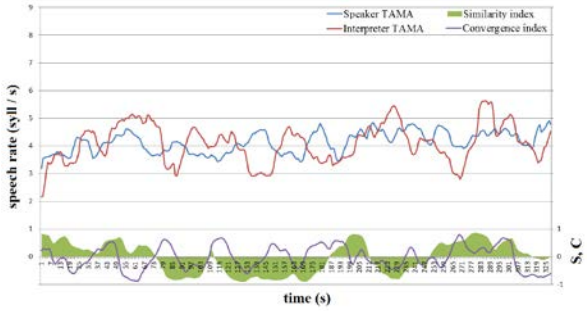


Figure 3: Speech rate TAMA plot with the interpreting 'chasing' the speaker but following their own pace (S4/I3)

5.4 Pitch range

In all samples, a high degree of synchrony in pitch range was observed (see Table 4). However, the interpreter's pitch range is smaller than the speaker's. This suggests that interpreters limit the extent to which they produce emphatic speech.

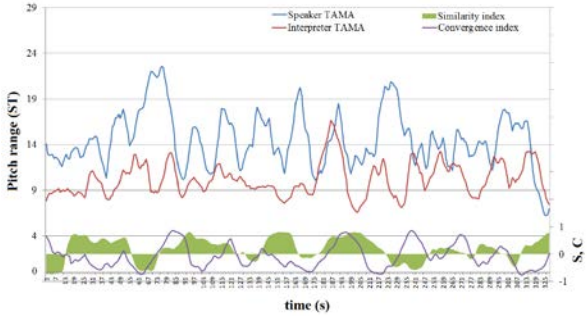


Figure 4: A typical pitch range TAMA plot

5.5. Mean pitch

Although the normalised time series are mostly similar (during 39% to 69% of the time), we do not observe convergence to the speaker's pitch, neither in absolute or relative terms. This seems logical since mean pitch imitation has been linked to mimicry in previous research (Couper-Kuhlen 1996).

6. Conclusions and further work

In conclusion, interpreters seem to adopt less expressive (more 'continuous') speaking style than the parliamentary speakers, a

manner that can be accounted for by the fact that their speech is dependent on, and subordinated to the speech of the orator. They seem to have their own strategies for responding to important changes in the speaker's prosodic cues, while at the same time exhibiting a distinct prosodic profile. The practical and cognitive constraints lead to a particular temporal organisation of their speech, with longer and less frequent silent pauses and a more variable speech rate.

We plan to expand this study, controlling for directionality (EN to FR, and FR to EN) and studying the relationship between the prosodic features and the structure of the original and the interpreters' speech.

Acknowledgments

I would like to express my gratitude to Prof. Anne-Catherine Simon for the guidance and encouragement she provided throughout this research.

References

Boersma, P. & Weenink, D. (2009). Praat: doing phonetics by computer. http://www.praat.org

Bigi, B. (2012). SPPAS: a tool for the phonetic segmentation of speech. *The 8th LREC*, Istanbul (Turkey), May 2012.

Cannam, C., Landone, C., Sandler M. (2010). Sonic Visualiser: An open source application for viewing, analysing, and annotating music audio files, *Proceedings of the ACM Multimedia 2010 International Conference*, pp. 1467-1468.

Couper-Kuhlen, E. (1996). The prosody of repetition: on quoting and mimicry. E. Couper-Kuhlen & M. Selting (eds.), *Prosody in conversation, studies in interactional sociolinguistics,* Cambridge University Press, pp. 366-405.

De Looze C., & Rauzy S. (2011). Measuring speakers' similarity in speech by means of prosodic cues: methods and potential. *Interspeech 2011*. pp. 1393-1396.

Eskenazi, M. (1993). Trends in speaking styles research, ISCA, pp. 501-509.

Gile, D. (2009). Basic concepts and models for interpreter and translator training (revised ed.) John Benjamins, Amsterdam/ Philadelphia.

Goldman, J.-Ph., Auchlin, A., Simon, A.C. & Avanzi, M. (2007). Phonostylographe: un outil de description prosodique. Comparaison du style radiophonique et lu. *Nouveaux cahiers de linguistique française* 28, pp. 219-237.

Goldman, J.-Ph. (2008). EasyAlign: a semi-automatic phonetic alignment tool under Praat, http://latlcui.unige.ch/phonetique

Goldman, J.-P., François, T., Roekhaut, S., Simon, A.-C. (2010): Étude statistique de la durée pausale dans différents styles de parole. Association Francophone de la Communication Parlée (ed.): *Actes des 28èmes Journées d'Étude sur la Parole. JEP 2010*. Mons, Belgium, 25-28 May 2010

Goldman, J.-Ph., Auchlin, A. & Simon, A.C. (2011). Description prosodique semi-automatique et discrimination des styles de parole. Yoo, H-Y & Delais-Roussarie, E. (eds.), *Actes d'IDP 2009*, Paris, Septembre 2009, ISSN 2114-7612, pp. 207-221.

Goldman, J.-Ph., Avanzi, M., Simon, A.C., Auchlin, A. (2012). A continuous prominence score based on acoustic features. *Interspeech 2012, 13th Annual Conference of the International Speech Communication Association*, Portland, Oregon, USA, 9-13 September 2012.

Kousidis, S., Dorran, D., Mcdonnell, C. and Coyle,E. (2009). Convergence in human dialogues. Time Sries Analysis of Acoustic Features, *Proceedings of SPECOM 2009*, St. Petersburg, p. 2.

Léon, P. (1993). Précis de phonostylistique, Parole et expressivité, Nathan Université, Paris.

Llisteri, J. (1992). Speaking styles in speech research, *ELSNET/ESCA/SALT Workshop on Integrating Speech and Natural Language*, Dublin, Ireland, p. 28

Mertens, P. (2004). The Prosogram: Semi-Automatic Transcription of Prosody based on a Tonal Perception Model. Bel, B. & Marlien, I. (eds.), *Proceedings of Speech Prosody 2004*, Nara (Japan), 23-26 March. ISBN 2-9518233-1-2.

Pöchhacker, F. (2004). Introducing interpreting studies. Routledge, London / New York.

Simon, A.-C., Avanzi, M., Goldman, J.-Ph. (2008). La détection des proéminences syllabiques. Un aller-retour entre l'annotation manuelle et le traitement automatique, *Congrès Mondial de Linguistique Française 2008*, p. 151, DOI: 10.1051/cmlf08256

Simon A.C., A. Auchlin, M. Avanzi & Goldman, J.-Ph.(2010). Les phonostyles: une description prosodique des styles de parole en français. Abecassis, M. & G. Ledegen (eds.), *Les voix des Français. En parlant, enécrivant*, Peter Lang, Berne, pp. 71-88.

Les variables temporelles dans le dialogue

Jean-Philippe Goldman[✈], *Antoine Auchlin*[✈], *Anne Catherine Simon*[✲]

jean-philippe.goldman@unige.ch, antoine.auchlin@unige.ch,
anne-catherine.simon@uclouvain.be
[✈]Département de Linguistique, Université de Genève
[✲]Institut Langage & Communication, Université de Louvain (Louvain-la-Neuve)

Abstract

Time and rhythmic variables of speech were initially studied by Grosjean & Deschamp (1972) in order to measure and compare temporal characteristics in different languages and speaking styles. Aside this approach dedicated to monologues, some studies applied notions from conversation analysis to dialogues corpus in an automatic way.

Our contribution extends this work in both directions: (i) we develop a semi-automatic tool for temporal variables description in dialogues and (ii) we analyze them in a corpus of 35 dialogues to show how they vary according to situational features.

1. Les variables prosodiques temporelles

Les variables temporelles de la parole ont fait l'objet de nombreuses études, dans le domaine de la synthèse en vue d'améliorer le naturel de la parole (Zellner 1998), pour la modélisation du dialogue homme-machine (Heldner & Edlund 2010), ou en analyse des conversations, selon une approche qualitative (Auer, Couper-Kuhlen & Frank 1999) ou quantitative relevant de la linguistique de corpus (ten Bosch, Oostdijk & Boves 2005). Enfin, d'autres études s'intéressent aux aspects temporels dans une visée descriptive (Duez 1987, Campione & Véronis 2002, Goldman, François, Roekhaut & Simon 2010), contrastive (Grosjean & Deschamp 1975, Schwab 2007) ou pour étudier des phénomènes connexes comme les hésitations (Candea 2000).

Ces études, nombreuses et portant sur un large éventail de langues et de situations de parole, fournissent des mesures souvent peu comparables à cause des procédures utilisées pour les extraire et les analyser. Citons deux points qui font débat : le choix des auteurs d'écarter ou non des pauses "micro", inférieures à un certain seuil de durée (Campione & Véronis 2002 pour une critique), ou celui de représenter la distribution des durées pausales en valeurs brutes ou log-transformées (Heldner & Edlung 2010: 561-2 privilégient la première option). Notre contribution visant à proposer un outil accessible pour mesurer automatiquement ces variables temporelles, nous commençons par rappeler quelques notions de base.

1.1. Dans les monologues

Après plus de quarante ans, le travail pionnier de Grosjean & Deschamp (1972) reste une référence pour définir les variables temporelles de la parole. Le temps total de *locution* se répartit entre le temps d'*articulation* (ou de phonation) et le temps de *pause*. Le rapport entre ces deux variables fournit un pourcentage indiquant le *taux d'articulation* du locuteur et "permet de mesurer le temps passé à articuler" (1972: 131). La *vitesse de parole* est le nombre de syllabes produites sur le temps de locution (habituellement en syllabes par seconde) tandis que la *vitesse d'articulation* ne tient compte que du temps d'articulation (hors pauses). Le temps de locution peut se décomposer selon le nombre et la longueur des *suites sonores* (séquence « de syllabes entre deux pauses non sonores ») et des pauses (1972: 131).

À ces variables primaires s'ajoutent des variables secondaires telles que les pauses

sonores (marques d'hésitation et syllabes allongées), les répétitions et les faux départs. Requérant une annotation manuelle fastidieuse, ces variables sont souvent ignorées par les études sur grands corpus.

1.2. Dans les dialogues

L'analyse de conversations plurilocuteurs complique sensiblement l'inventaire des variables temporelles par le fait que le canal de parole peut être occupé par plus d'un locuteur simultanément, ou par aucun, et que chaque événement (en particulier les pauses silencieuses) doit être attribué à un locuteur spécifique. La notion même de tour de parole, centrale à cet égard, est très complexe à implémenter dans un système d'analyse automatique, tant cette unité résulte d'une analyse dynamique de la manière dont les participants combinent les *turn constructional units* (TCU's) de manière incrémentale pour produire ce qui sera considéré, en contexte, comme un tour de parole (Selting 2000, Mondada 2008). L'annotation automatique doit privilégier la notion de production verbale (pv), suite de syllabes attribuées à un locuteur, à celle de tour de parole (Groupe Icor 2006).

Une production verbale correspond dans certains cas à un signal de backchannel, une production brève (typiquement *mh* en français) réalisée ou non pendant une pause, par laquelle l'interlocuteur indique au locuteur en cours qu'il l'écoute et que ce dernier peut poursuivre son tour.

Quant aux pauses silencieuses, elles peuvent se produire à l'intérieur de la production verbale d'un locuteur (pause intra-locuteur) ou entre la fin de la production verbale d'un locuteur et le début de celle du locuteur suivant (pause inter-locuteur ou *gap*).

Le passage d'une production verbale à l'autre peut se faire sans pause ni chevauchement (appelé *no-gap–no-overlap par* Sacks, Schegloff & Jefferson 1974), mais donne régulièrement lieu à un chevauchement de parole (*overlap*) qui, le plus souvent, n'est pas ressenti comme une interruption du locuteur en cours mais comme une transition anticipée (Jefferson 1983).

Le modèle le plus complet des schémas de changement de tour est fourni par Weilhammer & Rabold (2003) qui dénombrent 10 cas. Ce modèle a été souvent repris, mais de manière simplifiée (voir par ex. ten Bosch et al. 2005, Heldner & Edlund 2010) car son application demande une annotation manuelle de certains phénomènes, comme les backchannels.

2. Méthode

2.1. Données

Cette étude se base sur l'analyse de 35 extraits de dialogues[1] représentatifs de différentes activités de parole. Par exemple, on y trouve des interviews, des revues de presse radiophoniques, des commentaires sportifs ou des demandes d'itinéraire. Afin de rendre ces données comparables entre elles, nous les avons catégorisées selon des critères situationnels (Koch & Oesterreicher 2001) dont nous faisons l'hypothèse qu'ils influencent, seuls ou se combinant entre eux, les variables temporelles. Ces critères sont au nombre de trois: degré d'interactivité, de préparation et caractère médiatique[2]. Selon cette typologie, nos données se répartissent comme dans le Tableau 1.

[1] Ces extraits proviennent des corpus Phonogenres (Audrit et al. 2012), Valibel et Rhapsodie (Lacheret et al. 2013), ce dernier étant une banque de corpus annotés de provenances diverses.

[2] *Interactif*: la parole est librement distribuée; semi-interactif: liberté d'interrompre restreinte; non-interactif: pas de liberté d'interrompre. *Préparé*: texte lu; semi-préparé: sujet connu et phraséologie disponible. *Médiatique* : exclusivement produit pour être diffusé (bulletin de nouvelles), secondairement médiatique : activité (débat, interview) impliquant plusieurs rôles communicatifs, médiatisée (retransmise sur les ondes).

2.2. Annotation

Chaque fichier sonore est accompagné d'un alignement (semi-automatique) au niveau du phonème, de la syllabe et du mot.

	interactif (total = 14)	semi-inter. (total =16)	non-inter. (total = 5)
Spontané (total = 16)	1,4, **3**	3, 1	4
Semi-prép (total = 16)	2,4	5,5	
Préparé/lu (total = 3)		1,1	**1**

*Tableau 1. Répartition des échantillons de parole selon les métadonnées (données non médiatiques (total = 16), secondairement médiatiques (total = 15), **médiatiques** (total = 4)).*

Selon les corpus, ces alignements sont présentés dans *un fichier par enregistrement*, avec une couche d'annotation manuelle précisant quel locuteur a la parole, ou d'*un fichier par locuteur*. Afin d'uniformiser, le modèle *un fichier par enregistrement* a été privilégié, fusionnant les couches d'annotation le cas échéant.

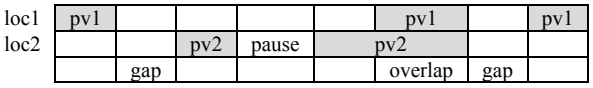


Figure 1. Occupation de la parole dans les dialogues.

Ces annotations permettent de catégoriser automatiquement ces phénomènes temporels (Figure 1) : productions verbales de chaque locuteur (*pv*); pauses intralocuteur (*pause*); pauses interlocuteurs (*gap*); chevauchement de parole (*overlap*). La distinction entre *pause* et *gap* devient problématique dès qu'un intervalle chevauché jouxte un silence. Edlund et al. (2009) la systématisent avec les notions d'*instigateur* et de *propriétaire* de la pause :

"*L'instigateur* d'un silence est le locuteur qui a parlé le dernier avant le silence (ou le dernier a avoir parlé seul, en cas d'arrêt simultané) ; le *propriétaire* du silence est celui qui le rompt (ou l'instigateur en cas de départ simultané) ; un *gap* est un silence dont l'instigateur n'est pas le propriétaire (silence inter-locuteurs) ; une *pause* est un silence dont l'instigateur est propriétaire (silence intra-locuteur)." (notre traduction).

2.3. Mesures

A partir des distinctions posées ci-dessus, nous proposons un outil qui renvoie les mesures suivantes, pour l'ensemble d'un enregistrement et par locuteur :

- Temps d'enregistrement
- Temps de locution (excluant les pauses liminaires et les passages inaudibles)
  - Temps d'articulation
    - articulation exclusive (d'un loc.)
    - chevauchement
    - articulation cumulée
    - production verbale
  - Temps de silence
    - pauses (intra)
    - gaps (inter)

Ces mesures simples sont exprimées en nombre (pauses, pv, tours de parole) et en durée (secondes). Des variables complexes sont dérivées, leur taux (% par rapport au temps de locution), leur durée moyenne et leur fréquence (par mn.), mesures également détaillées par locuteur. L'articulation cumulée additionne les temps d'articulation des locuteurs et peut donc être supérieure au temps de locution, et le taux d'articulation cumulée peut être supérieur à 100%. Les intervalles chevauchés sont comptés par locuteur, pointant le locuteur initiant le plus de chevauchements. Ces chevauchements peuvent donner lieu à une passation de tour (L1 → overlap → L2) ou non (L1 → L1+L2 → L1). Ce dernier cas peut être assimilé à un backchannel ou à une tentative de prise de tour sans suite.

En pratique, l'outil propose trois sorties différentes. Les deux premières affichent toutes ces mesures, l'une en synthétise une partie; la troisième crée un tableau. Dans les trois cas, on peut produire ces mesures pour plusieurs enregistrements et fusionner les résultats dans un seul tableau, à fin de comparaisons. La présentation synthétique des contributions de chaque locuteur sur un

extrait de 5 minutes se présente comme suit:

```
######Processing TextGrid foot0...
Tps de locution    299
  -spk 1             126   (42.1%)
    -art.exclus        79  (26.5%)
    -chevauchement     11  (3.5%)
    -pause             36  (12.1%)
  -spk 2             160   (53.2%)
    -art.exclus        124 (41.4%)
    -chevauchement     11  (3.5%)
    -pause             25  (8.3%)
  -gap               25    (8.2%)
```

3. Résultats et discussion

Une série de variables temporelles ont été mesurées pour l'ensemble du corpus puis contrastées selon deux traits situationnels.

3.1. Distribution globale des transitions de tours (pauses inter- et chevauchements)

Parmi les intervalles interlocuteurs, on distingue les chevauchements et les gaps, les premiers, (Fig. 2) en rouge (à gauche), représentés en « temps négatif » mesurant la fin de la *pv* du locuteur chevauché ; les *gaps*, en vert, constituent la majorité des cas de transition de tour avec un mode vers 250ms, distribution similaire à celle observée par Heldner & Edlung (2010 : 562) dans le Spoken Dutch Corpus. Comme eux, on n'observe pas de pic autour de 0 seconde; le cas *no-gap-no-overlap* n'est donc pas aussi fréquent que supposé par Sacks 1974.

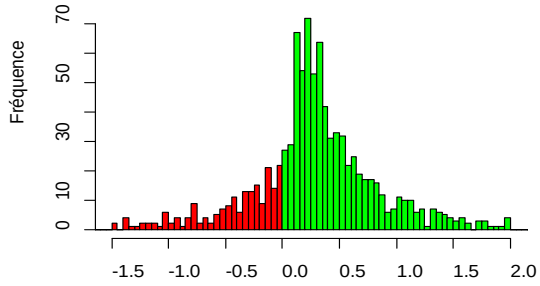


Figure 2. Distribution de la durée des 200 chevauchements (overlap) et des 752 gaps.

3.2. Comparaison - degré d'interactivité

La distribution des variables temporelles est-elle homogène dans le corpus, ou est-elle liée à un trait situationnel ? Notre première hypothèse était que ces variables seraient sensibles en premier lieu au degré d'interactivité.

Le Tableau 2 montre que si le degré d'interactivité joue un rôle minime dans le taux d'articulation – relativement constant –, il influence significativement la distribution des chevauchements (dont le taux augmente avec le degré d'interactivité[3]; $t(20)=-2.2, p<0.05$), la durée moyenne du tour (qui diminue quand l'interactivité augmente; $t(29)=3.2, p<0.01$) et la distribution globale des durées de tours ($t(30)=3.6, p<0.01$). Pour ces 3 mesures, les groupes d'échantillons semi- et non-interactifs semblant se comporter de la même manière ont été fusionnés pour les t-tests indiqués entre parenthèses.

Interactif	Non	Semi	Oui
Nb d'extraits	5	16	14
Temps de parole	102.3	279.0	332.1
Ratio d'articulation	79.4	80.8	77.3
Ratio d'overlap	**2.5**	**2.8**	**6.4**
Durée moy. de pv	**13.4**	**13.5**	**5.8**
Dur. moy. de pv loc1	**21.5**	**22.1**	**8.8**
Ratio du loc.1	90.2	84.6	65.5
Durée moy.des gaps	0.5	0.8	0.6
Ratio de gaps	5.1	3.8	7.4

Tableau 2. Distribution des mesures temporelles selon le degré d'interactivité.

3.3. Comparaison - degré de préparation

Le degré de préparation (spontané, semi-préparé, préparé) joue aussi un rôle important dans l'organisation temporelle des dialogues. Comme le montre le Tableau 3, la durée des tours augmente avec le degré de préparation du discours, de même que le temps de parole du locuteur principal de manière significative. Le taux de chevauchement, quant à lui, est plus élevé dans le discours plus spontané.

[3] Rappelons que notre corpus est dialogal; le discours non interactif peut contenir des chevauchements.

Préparation	Spont(16)	Semi(16)	Prép(3)
Temps de parole	242.0	321.0	205.4
Ratio articulation	77.8	81.1	77.0
Ratio d'overlap	**5.4**	**3.6**	**0.4**
Durée moy. de pv	**8.1**	**10.0**	**24.4**
Dur.moy.de pv L1	**12.8**	**17.0**	**35.4**
Ratio du L1	74.9	**79.4**	**83.8**
Dur.moy.des gaps	0.7	0.7	1.0
Ratio de gaps	6.7	4.6	2.6

Tableau 3. Distribution des mesures temporelles selon le degré de préparation de la parole

4. Conclusion

L'article montre que les variables temporelles dans les dialogues sont sensibles aux traits situationnels tels que le degré d'interactivité ou de préparation du discours. Il permet de tirer des conclusions plus générales et généralisables que celles issues d'études contrastant deux styles de parole (par ex. conversation téléphonique vs. en face à face, dans ten Bosch et al. 2005).

Remerciements

Ce travail a en partie été rendu possible grâce au projet FNS Suisse « Caractérisation linguistique et prosodique de styles de parole, approche semi-automatique et applications » (fonds n°100012_134818).

References

Audrit S., Prsir T., Auchlin A. & J.Ph. Goldman (2012). Sport in the media: a contrasted study of three sport live media reports with semi-automatic tools, *Proc. of Speech Prosody 2012*, Shanghai.

Auer, P., E. Couper-Kuhlen. & F. Müller (1999). *Language in Time. The Rhythm and Tempo of Spoken Interaction*. Cambridge University Press.

Campione, E. & J. Véronis. (2002). A Large-Scale Multilingual Study of Silent Pause Duration. *Proceedings of the Speech Prosody 2002 Conference*. Aix-en-Provence, pp. 199‑202.

Candea, M. (2000). *Contribution à l'étude des pauses silencieuses et des phénomènes « d'hésitation » en français oral spontané.* Thèse de doctorat. Université Paris III-La Sorbonne Nouvelle.

Duez, D. (1987). *Contribution à l'étude de la structuration temporelle de la parole en français*. Thèse de Doctorat d'Etat. Aix-Marseille 1.

Edlund, J., M. Heldner & J. Hirschberg (2009). Pause and gap length in face-to-face interaction. *Proceedings of INTERSPEECH*, pp. 2779–2782.

Goldman, J.-P., T. François, S. Roekhaut & A.C. Simon (2010). Étude statistique de la durée pausale dans différents styles de parole. *Actes des XXVIIIèmes Journées d'Etude sur la Parole*, Mons, 25 - 28 mai 2010. pp. 161‑164.

Grosjean, F. & A. Deschamps (1972). Analyse des variables temporelles du français spontané. *Phonetica*, 26, pp. 129‑156.

Grosjean, F. & A. Deschamps (1975). Analyse contrastive des variables temporelles de l'anglais et du français: vitesse de parole et variables composantes: phénomènes d'hésitation. *Phonetica*, 31, pp.144‑184.

Groupe ICOR. (2006). Glossaire. Site CORINTE http://icar.univ-lyon2.fr/projets/corinte/

Heldner, M. & J. Edlund J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4), pp.555–568.

Jefferson, G. (1983). Notes on some orderliness of overlap onset. *Tilburg Papers in Language and Literature* 28, Department of Linguistics, Tilburg University

Koch, P. & W. Oesterreicher. (2001). Langage parlé et langage écrit. In G. Holtus, M. Metzeltin, & C. Schmitt (éd). *Lexikon der romanistischen Linguistik*, I/2, pp.584‑627.

Lacheret A., S. Kahane & P. Pietrandrea (eds.). (2013), *Rhapsodie: a Prosodic and Syntactic Treebank for Spoken French*, Amsterdam, Benjamins.

Mondada, L. (2008). L'interprétation *online* par les co-participants de la structuration du tour *in fieri* en TCUs: évidences multimodales. *Tranel* 48.

Sacks, H., E. A. Schegloff & G. Jefferson. (1974). A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, 50(4), pp. 696‑735.

Schwab, S. (2007). *Les variables temporelles dans la production et la perception de la parole*. Thèse de Doctorat, Université de Genève.

Selting, M. (2000). The Construction of Units in Conversational Talk. *Language in Society*, 29/4, pp. 477‑517.

Simon, A.C., A. Auchlin, M. Avanzi & J.-Ph. Goldman (2009). Les phonostyles: une description prosodique des styles de parole en français. Abecassi, M. & G. Ledegen (éd.), *Les voix des Français. En parlant, en écrivant*, Peter Lang: Berne, 71-88.

Ten Bosch, L., N. Oostdijk & L. Boves (2005). On temporal aspects of turn taking in conversational dialogues. *Speech Communication*, 47, pp. 80‑86.

Weilhammer, K. & S. Rabold (2003). Durational aspects in turn taking. *International Congresses*

of Phonetic Sciences.
Zellner, B. (1998). *Caractérisation et prédiction du débit de parole en français. Une étude de cas.* Thèse de doctorat. Université de Lausanne.

Multifunctionality of prosodic boundaries in spontaneous narratives in Kammu

Anastasia M Karlsson[1], Jan-Olof Svantesson[1], David House[2]

anastasia.karlsson@ling.lu.se, jan-olof.svantesson@ling.lu.se, davidh@speech.kth.se
[1]Dept. of Linguistics and Phonetics, Lund University
[2]Dept. of Speech, Music and Hearing, KTH, Stockholm

Abstract

The main function of sentence intonation in Kammu is to mark prosodic boundaries. There is no additional tonal marking of focus. It is of particular interest that the underlying intonation system is the same for both tonal (Northern Kammu) and non-tonal (Eastern Kammu) dialects. Prosodic boundaries in Kammu have three functions: they mark prosodic phrases, focus and speaker engagement. In this study we show that relationships between boundaries in terms of upstepping or its absence interact with information and discourse structure. This relationship has the same pattern in both tonal and non-tonal Kammu.

1. Introduction

1.1. Kammu language

Kammu is a Mon-Khmer language. It is spoken by some 600,000 people mainly in Northern Laos, but also in adjacent areas of Vietnam and Thailand. One of the main dialects of this language is a tone language of the 'East Asian' type with (high or low) tone on each syllable, while the other main dialect lacks lexical tones.

The origin of the tones of the tonal dialect is due to the development of high pitch in vowels following a voiceless consonant and low pitch in vowels following a voiced consonant, and the subsequent merger of voiceless and voiced consonants into the unmarked member of the pair, voiceless for stops and voiced for sonorants. Thus, *puuc* 'to undress' became *púuc* (high tone) in the tonal dialect and *buuc* 'wine' became *pùuc* (low tone). The non-tonal dialect kept the original forms unchanged. Other differences, phonological, morphological or syntactic, between the dialects are marginal, and speakers of different dialects understand each other without difficulty (Svantesson 1983; Svantesson & House 2006).

1.2. Kammu intonation

The main function of sentence intonation in Kammu is to mark prosodic boundaries. Phrase boundaries occur at the right edge of each prosodic phrase and are realised by a high (or high falling) pitch. The focused word is by default placed at the end of an utterance coinciding with the place of the boundary tone, and the pitch of the phrase boundary tone is raised. There is thus no additional tonal gesture for focal accent. In the tonal dialect lexical tones do not change the phrase pattern, and we still find the high boundary tone at the right edge of prosodic groups unless it jeopardises the identity of the lexical tones (Karlsson et al. 2012).

2. Research questions, methodology, material

2.1. Research question and methodological framework

In Kammu, phrase boundaries between utterances said in isolation tend to be upstepped. Informal observations of spontaneous narratives indicate that besides upstepping of phrase boundaries within an utterance there is also upstepping between boundaries of utterances. The upstepping occurs up to a certain point and then the same

pattern repeats again. These turning points seem to occur at thematically similar places in narratives for all our speakers. Our goal is to find out whether these turning points are related to discourse structure. The main assumption is that tonal phrase boundaries in Kammu are multifunctional. They reflect prosodic phrasing, information structure and discourse structure. First, we assume that information structure is reflected by upstepping of phrase boundaries. The utterance final boundary is the highest one as a reflection of default placement of the focused word (or 'new' information) at the end of an utterance. Second, we assume that the long-term relations between utterance boundaries reflect discourse structure.

In our analysis we distinguish between information structure and discourse structure. Narratives are divided into [given + new] units, called major phrases. As the information becomes given a new major phrase starts. Each major phrase consists of at least one minor phrase. Minor phrases are defined on prosodic grounds. We recognise a group as a minor phrase if it has a prosodic boundary (high or high falling pitch) at its right edge. Discourse topics are recognised on semantic grounds; this is described in 3.2. The F0 contour of a part of a narrative with its division into minor and major phrases and topics is shown in Figure 1.

2.2. Material

Recordings of four speakers (all men) of the non-tonal dialect and six speakers (four women and two men) of the tonal dialect of Kammu were used for this investigation. They recorded spontaneous accounts of rice growing, from the beginning of the work in the field until the rice is cooked and eaten. All speakers are well acquainted with this process and their accounts are very similar. Thus, we got fairly homogeneous spoken texts lasting about 2–5 minutes each. The narratives were transcribed and glossed by a native speaker of Kammu.

3. Analysis

3.1. Informational structuring of the narratives

Structuring of new and old information is achieved in the same way by all speakers: new information is placed at the end of the utterance; it is then repeated in the next utterance and is followed by new information. The informational structuring is [anchor + new_1] [old new_1 + new $_2$] [old new_2 + new_3]… The new information becomes an anchor point (old information) in the next utterance. There are thus a lot of repeated words in the speakers' monologues, and anaphoric reference is not used. An example is (only key events are included):

Before there is rice we have to clear the field… After clearing the field we burn the field… After burning we sow.[1]

The text can thus be seen as a list of successive events. Some speakers use only one utterance per event while some add a lot of additional information. In order to test our hypothesis that each informational unit [given + new], i.e. major phrase, is reflected in the tonal structure by upstepping of boundary tones, we made two kinds of analysis. First, we divided narratives into phrases on prosodic grounds. This was done by perceptual and visual analysis of the F0 contours using *Praat*. Each unit ending with a prosodic boundary tone was labelled as a minor phrase and the F0 maximum on the last word was measured. Second, we performed an analysis of the informational flow in narratives in terms of 'new' and 'given'. Each unit consisting of 'given + new' is labelled as a major phrase. Thus, the division is: [[minor phrase]$_{boundary1}$[minor phrase]$_{boundary2}$[minor phrase]$_{boundary3}$]major phrase]$_{boundary4}$.

[1] Kammu people practice slash-and-burn agriculture.

Each major phrase consists of at least one minor phrase. The boundary of the last minor phrase is also the boundary of the whole major phrase. We expect the F0 maxima of boundary tones to be upstepped with the highest F0 at the boundary of the major phrase (boundary 4 in the example above).

3.2. Division of narratives into topics

A discourse topic is seen as an informatively coherent part of discourse with a clear beginning and end (see e.g. Chafe 2003). As we are dealing with narratives about a traditional activity we used the Kammu agricultural calendar compiled by Damrong Tayanin:
http://digaaa.humlab.lu.se/digaaa/web/kammu/KamRaw/kammu1.html.

The agricultural periods are:

1) Clearing, 2) First burning, 3) Second burning, 4) Sowing, 5) First weeding, 6) Second weeding, 7) Third weeding, 8) Harvest, 9) Finishing off the year, 10) Cold season.

Having these as our reference frame for division into thematic topics we found that all speakers have the following topics:

1) Clearing, 2) First and second burning, 3) Sowing, 4) Weeding, 5) Ripe rice, 6) Harvest, 7) Putting in barns, 8) Pounding rice, 9) Soaking rice, 10) Cooking rice, 11) Eating rice.

Some speakers have additional topics, such as making field houses, protecting crops from animals or different ways to cook rice. We chose only topics that were found for most speakers: all topics except (5) and (7) occur for all speakers.

As phrase boundary tones in Kammu convey several functions, we assume that they also interplay with discourse structure. As we observe upstepping of boundary tones within major phrases as a cue for their boundaries, we assume that also the end of topics will be marked by a higher boundary.

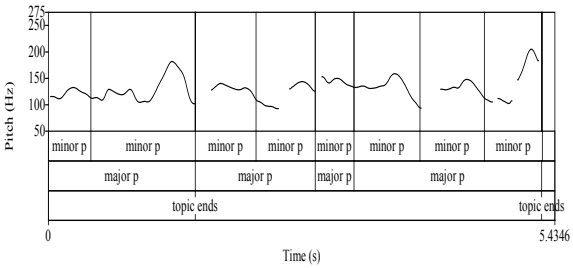


Figure 1. Part of a narrative and its division into minor phrases, major phrases and topics. Non-tonal speaker. Glossing is [[go mark]$_{minor\ phrase}$ [we finish then clear]$_{minor\ phrase}$]$_{major\ phrase}$]$_{topic\ ends}$ [[clear]$_{minor\ phrase}$ [finish then dry]$_{minor\ phrase}$]$_{major\ phrase}$ [[cut tree]$_{minor\ phrase}$]$_{major\ phrase}$ [[that one month]$_{minor\ phrase}$ [two months]$_{minor\ phrase}$ [finish then burn]$_{minor\ phrase}$]$_{major\ phrase}$]$_{topic\ ends}$.

4. Results

4.1. Final boundary tones of major-phrases

In order to find out whether the general pattern is that F0 is rising in major phrases, we measured the F0 maximum of the last word of each major phrase and of each minor phrase within the major phrases. As a measure of the F0 rise within a major phrase we took the difference between F0 of the major phrase and the mean of the F0 values of the (non-final) minor phrases that constitute the major phrase. For each speaker we thus obtained a number of differences which should be positive if the hypothesis that F0 increases in a major phrase is true. To test this hypothesis we used an exact binomial test for each speaker based on the number of positive and negative differences. For tonal speakers, the influence of the tones was compensated for by adding the mean F0 difference between the high and low tone in the measured words for that speaker to the F0 value of the maximum measured in each word with low lexical tone. The results are shown in Table 1. The tests show significant results (on the 5% level) for all speakers except Speaker 7 (non-tonal) and Speaker 18 (tonal), thus supporting our hypotheses for most speakers.

Non-tonal speakers:

Speaker	#Diff.=0	#Diff.>0	#Diff.<0	*p*-value
1	0	15	0	<0.001
6	1	29	1	<0.001
7	0	9	3	0.07
8	0	20	4	<0.001

Tonal speakers:

Speaker	#Diff.=0	#Diff.>0	#Diff.<0	*p*-value
17	0	11	1	0.003
18	0	10	4	0.09
19	0	9	2	0.033
20	0	12	1	0.0017
21	0	9	1	0.01
26	1	5	0	0.03

Table 1. The number of major phrases for which the difference between the F0 maximum of the major phrase and the mean F0 maxima of the constituent minor phrases is equal to, greater than or less than zero.

4.2. Boundaries of discourse topics

We tried to correlate the phrasing of the discourse with the local F0 maxima of the major phrases. In general there seems to be a tendency that local F0 maxima also serve as boundaries between discourse topics (in about 58% of the cases), but this is not always the case. The general trend of F0 maxima of boundary tones coinciding with the end of each topic is shown in Figure 2.

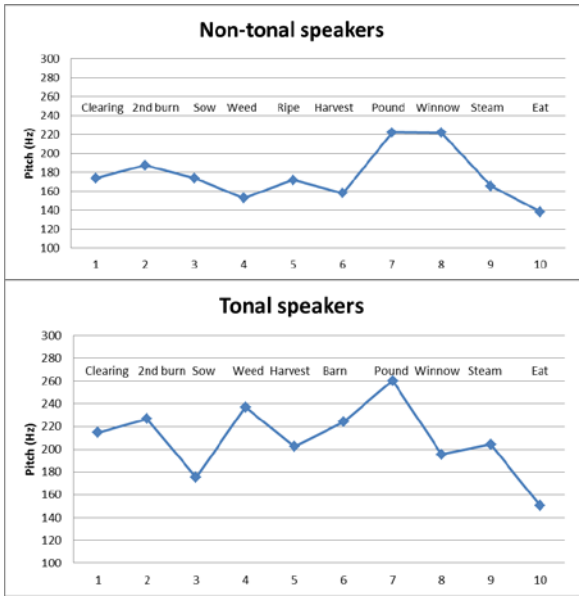


Figure 2. Mean of F0 maxima of topics

All speakers mark 'pounding' with the highest F0. After this point the general upstepping trend becomes opposite and we find downstepping between topics and also between boundaries of major phrases.

5. Discussion

5.1. Final boundary tones of major phrases

Spontaneous discourse encompasses many factors that may influence tonal patterns, such as phrasing, focusing, turn-taking, speakers' attitudes and degrees of engagement, self-corrections, hesitations, etc. Having investigated only one of these factors – topic marking – in our study, we have to keep in mind that our result may be influenced by all these factors. We chose to separate information structure and discourse structure in our study, which proved to be fruitful. Material was analysed by using three principles: prosodic analysis to extract tonal boundaries, analysis of informational status to detect major phrases and semantic analysis to decide topics. The three analyses were performed independently of each other and were then matched to see if our hypotheses are correct.

The division of narratives into information units [given + new] (major phrases) is reflected in prosodic phrasing by upstepping of boundary tones. We obtained statistical significance for both tonal and non-tonal speakers. As we move to discourse structure we can only talk about trends. All speakers mark the topic about pounding with the highest tonal boundary. Kammu speakers may see activities connected to rice as divided into two main parts: field work and cooking rice. Field work ends when one can pound the harvested and dried rice. Pounding is then the end of the first part of the narratives and is also marked by the highest boundary.

The end of other topics tends also to be tonally marked by a higher boundary. This trend is, however, broken by two main

factors. The part after 'pounding', in which the rice is cooked and eaten, shows the opposite trend: boundaries of all units (both of major phrases and topics) tend to decline. Thus the end of discourse is marked by a long-term downtrend of prosodic boundaries.

Due to the character of the structure of the narratives, we assumed that topics are structured as [description + name of activity when it is finished]. For example, in

We go to seek a field, seek in the forest, after finding the field we clear

the part before *clear* will be the description, and *clear* is the name of the activity and its ending, coinciding with the end of the topic. However, in some cases we found another type of structuring of topics, when the topic is introduced at the beginning and then described, e.g.:

We seek a place we will clear, yes, seek the forest, look for a place that will be good for the rice and we clear.

Here, *clear* is introduced in the beginning as a new topic and its development comes afterwards. This kind of topic gets the highest F0 at the beginning of the topic instead of at the end.

5.2. Typological implications

As regards prosodic typology, Kammu belongs to the phrase language type in Féry's (2010) typology. In this type of language, information structure is most often conveyed by morpho-syntactic means, and focusing is achieved by changes in the pitch level of phrasing tones, dephrasing or insertion of a new boundary tone. No new pitch accents are added to mark a focused word as is the case in intonation languages. According to this description, major Indian languages as Hindi, Bengali, Tamil and Malayalam (Féry 2010), as well as Korean (Jun 2005), West Greenlandic (Arnhold, to appear) and Mongolian (Karlsson, to appear) are typical phrase languages. Kammu has one type of boundary tone realised with a high (or high falling) pitch. Boundaries are multifunctional and they convey phrasing, focus, engagement, and topic structure. The occurrence of lexical tones does not lead to any differences, and we find the same strategies in conveying discourse structuring into topics in both tonal and non-tonal speakers.

Acknowledgements

The work reported here was done within the project *Integrating the structures of information and discourse: a cross-linguistic approach,* financed by the Swedish Research Council.

References

Arnhold, A. (to appear). Prosodic structure and focus realization in West Greenlandic. Jun, S.-A. (ed.), *Prosodic Typology Volume II*. Oxford University Press, Oxford. To appear.

Chafe, W. (2003). The analysis of discourse flow. Schiffrin, D., D. Tannen & H. E. Hamilton (eds.), *The handbook of discourse analysis*. Blackwell Publishing. Blackwell Reference Online.

Féry, C. (2010). Indian languages as intonational 'phrase languages'. Hasnain, S. I. & S. Chaudhury (eds.), *Problematizing language studies: cultural, theoretical and applied perspectives – essays in honor of Rama Kant Agnihotri.* Aakar Books, Delhi, pp. 288–312

Jun, S.-A. (2005). Prosodic Typology. Jun, S.-A. (ed.), *Prosodic typology: the phonology of intonation and phrasing*. Oxford University Press, Oxford, pp. 430–458.

Karlsson, A., D. House & J.-O. Svantesson (2012). Intonation adapts to lexical tone: the case of Kammu. *Phonetica* 69, pp. 28–47.

Karlsson, A. (to appear). Intonation in Halh Mongolian. Jun, S.-A. (ed.) *Prosodic Typology Volume II.* Oxford University Press, Oxford.

Svantesson, J.-O. (1983). *Kammu phonology and morphology*. Gleerup, Lund.

Svantesson, J.-O. & D. House (2006). Tone production, tone perception and Kammu tonogenesis. *Phonology* 23, pp. 309–333.

Temporal patterns of segments and intervals in Hungarian language

Anna Kohári

kohariannna@gmail.com
Department of Phonetics, Eötvös Loránd University

Abstract

Understanding the general trends and variability in the temporal patterns of speech has attracted attention over many years. In this preparatory work novel techniques and mathematical tools are presented to quantify timing structures within utterances in read and spontaneous speech. Statistical tests were performed addressing the frequency and magnitude distributions of decelerating and accelerating durational patterns of Hungarian speech. The proposed parameters were found to be sensitive indicators of the general asymmetry under time-reversal that is present in the timing dynamics of speech.

1. Introduction

The temporal patterns of speech exhibit far greater structural complexity and variability than what could be explained solely by the individual properties of the segments involved. These patterns are present in the form of lengthening and shortening effects, which influence local speech rate to a large extent. The present work attempts to analyze the statistical properties of these phenomena using methods that have rarely been used in phonetic research.

As a general tendency, it is well known that lengthening may often occur in the utterance final position in various languages (see Wightman et al. 1992). For example, in Hungarian language, vowels were found to be longer in utterance-final position than in utterance-medial position (see e.g. White & Mády 2008). Another basic phenomenon is the so-called polysyllabic shortening, which means that the durations of segments within a word tend to decrease as the length of the word increases (see e.g. Lehiste 1972). In Hungarian, this tendency can be observed in stressed and unstressed positions as well (Tarnóczy 1974). Obviously, many further effects may also lead to lengthening or shortening, but for Hungarian language only these two long-range tendencies could be conclusively identified (Gósy 2004).

Another possible approach to the understanding of these temporal patterns can be taken using the framework of speech rhythm. In a certain speech rhythm model it has been verified that using metrics based on the so-called vocalic and consonantal intervals, the timing characteristics of different languages can be determined (Ramus et al. 1999, Grabe & Low 2002). One class of these rhythm metrics (the Pairwise Variability Indices or PVIs) measures the variability of duration differences between consecutive intervals. The fact that PVIs have proven useful for describing speech timing has driven attention towards measures based on consecutive intervals (Grabe & Low 2002). These metrics, however, have mainly been applied to capture the general properties of utterance-scale or larger units, whereas the temporal variability within utterances has barely been investigated.

The scope of the present work was to apply this methodological framework to obtain an appropriate quantification of acceleration and deceleration within utterances. Therefore, the magnitudes and frequencies of accelerating and decelerating blocks were determined, and the length statistics of such intervals were analyzed.

2. Methods

The analyses in the present work were based on a corpus of spoken Hungarian, the BEA audio database (Gósy 2008). Recordings of 10 native speakers (5 males, 5 females; 20-60 years) reading out the same set of 20 simple sentences were investigated. Similarly, 20 utterances from the spontaneous speech of each of the same 10 participants were also evaluated.

Semi-automated phonetic segmentation of these records has been performed by the Munich AUtomatic Segmentation (MAUS) software (Schiel 1999) and manually corrected, applying the Praat 5.1 segmentation software using segmentation criteria similar to those used by Grabe & Low (2002).

C++ scripts were implemented to produce further numerical parameters from the 'raw' segmental duration series. Firstly, the successive vowels and consonants were merged into compact vocalic (V) and consonantal (C) intervals. By definition, this practice leads to alternating ...CVCV... structures. In the subsequent investigations the utterances were analyzed in parallel based on the following subsets of data: segmental level (taking segmental durations as temporal units); V intervals only (V intervals as units, ignoring C intervals); C intervals only (C intervals as units, ignoring V intervals); CV intervals (taking merged intervals as units, regardless of their 'identity'). Pauses and hesitations were excluded, as by Grabe & Low (2002).

Statistical tests were applied to quantify the time-reversal asymmetry of the duration records. The tests were conducted as follows. Firstly, the forward differences of the durations of adjacent units (segments or intervals) were calculated along the time series (i.e. always the duration of the successive unit is subtracted from that of a given unit). The number of such increasing (decreasing) steps of the record, when the forward difference was positive (negative), were logged, and denoted by N_I (N_D). Similarly, the average magnitudes of increasing and decreasing steps (marked $\langle\Delta t_I\rangle$ and $\langle\Delta t_D\rangle$) were measured. The step number ratios N_I / N_D and average step size ratios $\langle\Delta t_I\rangle / \langle\Delta t_D\rangle$ quantified the extent of asymmetry under time reversal (Gyüre et al. 2007).

Fig.1a depicts the parameter plane of the above two metrics. If both ratios were exactly 1, that would imply perfect temporal symmetry. If the time series was asymmetric but stationary (as curve A of panel Fig.1b), the data point would lie on the thick black curve, that corresponds to inverse relationship between the two quantities, hereafter referred to as the 'stationary curve'. Then the small but numerous increasing steps would be balanced by less frequent but proportionally larger decreasing steps, as in the case of A. Curve B is from a non-stationary region where the larger duration decrements (i.e. accelerations) are overruled by the smaller but more frequent duration increments (decelerations). If both ratios exceeded unity, the increasing steps would 'win' concerning their frequencies and magnitudes as well (curve C).

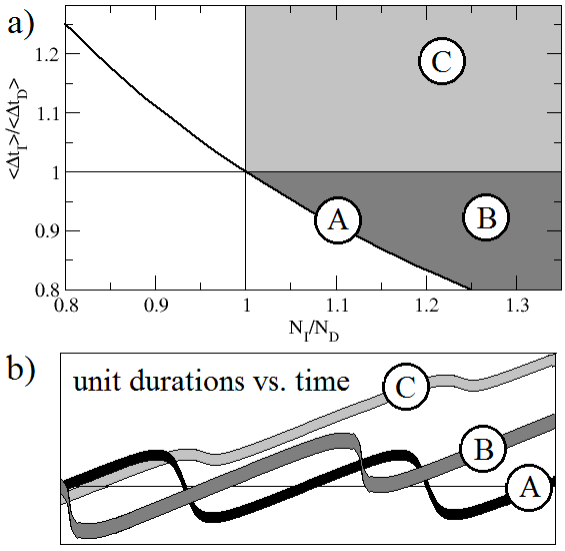


Figure 1: schematic drawing of the parameter plane of step number ratios and average step size ratios (a), and four sketches of time series (b) at the marked locations on the plane (A,B and C).

As a Monte Carlo significance test of the results, the iterative Fourier surrogate 'data shuffling' method of Schreiber and Schmitz (2000) and their open source Time Series Analysis (TiSeAn) 3.0.1 package was applied. For each real time series 10 random shuffled test series were created.

The length distributions of accelerating and decelerating blocks in the duration time series were also evaluated. An accelerating (decelerating) block is defined as a compact block of units in which the duration of a unit is always larger (smaller) than that of its successive neighbour. The length of such a block can be expressed as the number of the units it consists of.

3. Results

3.1. Step number and step size ratios

Correlation plots between step number and average step size ratios are presented in Fig.2 for the different types of basic units (segmental, V, C and CV) and speech styles (spontaneous and read). The parameter values calculated for the 'shuffled' time series are also shown for comparison (orange and turquoise dots).

The general tendency to be observed in the data is that the frequency of increasing steps was markedly larger than that of the decreasing steps, i.e. the majority of duration differences between the successive units exhibited positive values, regardless of the basic units used. Moreover, the fact that almost every measured data point in Figs. 2a-d is located above the dashed stationary curve, indicates an overall increasing trend of durations within the utterances (cf. Fig.1). Note, that the data points of the shuffled time series are scattered along the stationary curve and show symmetric distribution around unity in terms of both variables. In all four panels, the actual data points form a clearly distinct cluster from their shuffled counterparts, which underlines the robustness of the aforementioned trends compared to statistical fluctuations.

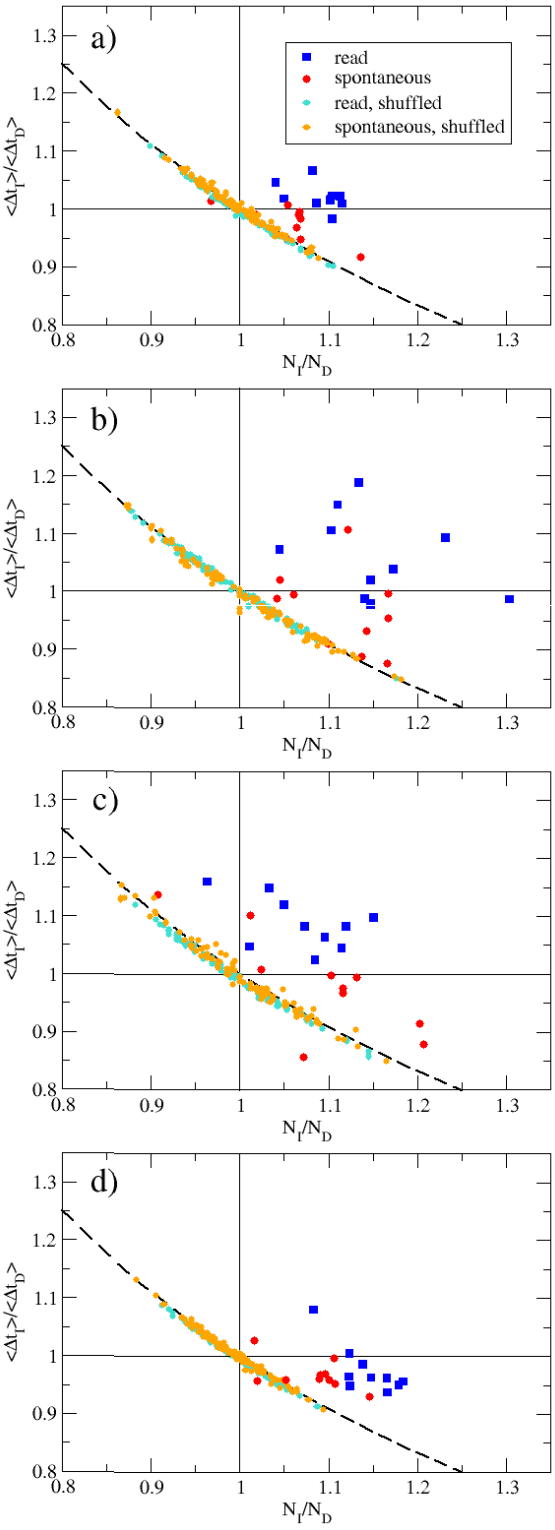


Fig.2: Correlation plots between step number ratios and average step size ratios for the cases where segments (a), vocalic intervals (b), consonantal intervals (c), and general intervals (d) were used as duration units. See also the legend in panel a).

As the differences between the studied speech styles are concerned, it was found that the geometric centers of the 'read' groups (blue squares in Figs. 2a-d) are shifted to larger distances from the origin (and from the stationary curve) on the parameter plane compared to the 'spontaneous' results (red circles in Figs. 2a-d). Fig.2a shows the case in which phonetic segments were treated as basic units. Except for the spontaneous production of one speaker, the step number ratios were found to be greater than one.

As an interesting qualitative difference between speech styles, one can notice that the 'read' data points are scattered above the unit step size ratio, whereas for the 'spontaneous' ones the decreasing steps have larger average magnitude. Yet, the latter is compensated by the large step number ratio, leading to utterance deceleration. Fig.2b depicts the results for the case where the basic units were vocalic intervals (V). The overall structure of the scatter plot is similar to the previous case. It is to be noted, however, that the values of the parameters are distributed in a larger area of the plane, indicating that V intervals can serve as the basis of a sensitive measure of temporal asymmetry. Considering consonantal intervals (C) as basic units (Fig. 2c), the 'read' and 'spontaneous' data points formed markedly disjoint clusters in terms of both parameters.

For the case in which both C and V intervals were treated as units (regardless of their identity), the most apparent feature of the scatter plot (Fig.2d) is that the variability of the parameter values is markedly smaller than either for the C or V intervals. This feature is explained by the fact that here the unit intervals are 'physically' adjacent to each other (as in the case of segmental units, cf. Fig.2a), and thus the overall decelerating trend manifests itself in smaller values of $\langle \Delta t_I \rangle$.

3.2. Block length distributions

The length distributions of the decelerating blocks (continuously increasing unit durations) and accelerating blocks (decreasing unit durations) are shown in Figs. 3a and b, respectively. The color coding indicates the 'block length', i.e. the number of units that made up the blocks.

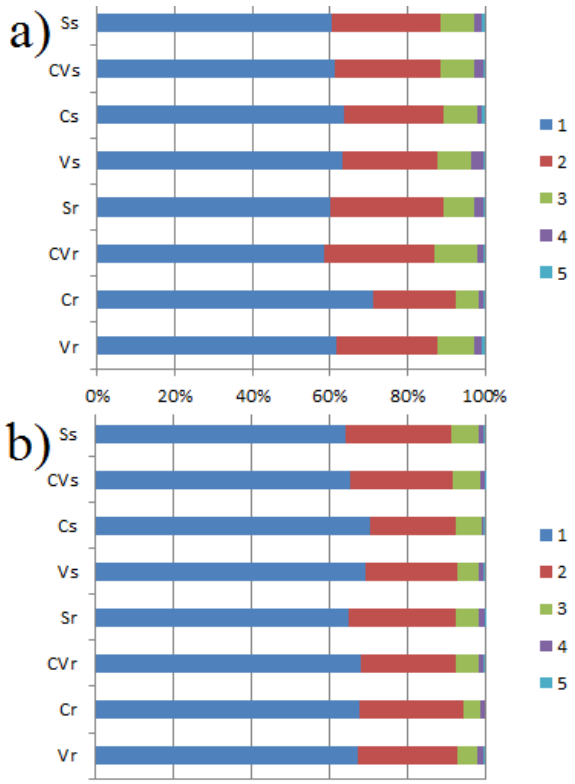


Fig.3.: Length distribution charts of decelerating (a) and accelerating (b) blocks. For the corresponding length numbers, see the color key. The abbreviations used: S - segmental, CV - general interval; C - consonantal interval; V - vocalic interval; s - spontaneous; r - read.

The relative frequencies of the block lengths are presented for different speech styles (read or spontaneous) and unit types (segmental, V, C or CV intervals) in both subfigures. It is visible that all these groups show highly similar distributions.

The next important remark is that (62.5±2.5)% of the blocks are actually single units. Note, however, that if one

multiplies the relative frequencies from these charts by their corresponding block length values it becomes clear that more than 50% of the units are actually members of a block of length 2 or more. This clearly implies clusterization in the unit duration differences.

4. Discussion

The utterances in spoken Hungarian were found to show a general trend of deceleration that could be properly quantified using novel statistical tools. This is consistent with recent measurements (Váradi & Beke 2013), and also with the general practice in speech synthesis that sentences are usually produced with decreasing speech rate (Olaszy 2006). The effects that yield deceleration (e.g. utterance-final lengthening) overcome the factors that lead to acceleration (e.g. polysyllabic shortening), both in frequency and magnitude.

The analysis of length distributions of compact accelerating and decelerating blocks indicate that these duration changes emerge in a fluctuating manner, instead of forming gradually increasing or decreasing, easily predictable patterns. This implies that the factors (e.g. utterance-final lengthening, focus, stress) which may modify the duration patterns within the utterances affect few subsequent units only. The relevance of the rarely appearing longer decelerating and accelerating blocks is that they may reveal the key locations where the speakers' intentions or emotional state are expressed.

5. Conclusions

The applied parameters were found to be useful indicators of time-reversal asymmetries of the timing dynamics of Hungarian speech. Further research is intended to clarify how the phenomena revealed here are realized in other languages.

Acknowledgments

I thank Miklós Vincze for his help in the numerical evaluation of the results. I am grateful to Mária Gósy and Alexandra Markó for the fruitful discussions and for making this research possible.

References

Boersma, P. & D. Weenink (2009). *Praat: doing phonetics by computer*. http://www.praat.org/

Gósy, M. (2004). *Fonetika, a Beszéd Tudománya.* [*Phonetics, the Science of Speaking*]. Osiris, Budapest.

Gósy, M. (2008). Magyar spontánbeszéd-adatbázis – BEA. [Hungarian Spontaneous Speech Database]. Gósy, M. (ed.) *Beszédkutatás 2008*, pp. 194-207.

Grabe, E. & E. L. Low (2002). Durational variability in speech and the rhythm class hypothesis. *Papers in Laboratory Phonology* 7, pp. 515-546.

Gyüre, B., I. Bartos & I.M. Jánosi (2007). Nonlinear statistics of daily temperature fluctuations reproduced in a laboratory experiment. *Physical Review E* 76, 037301.

Lehiste, I. (1972). The timing of utterances and linguistic boundaries. *The Journal of the Acoustical Society of America* 51, pp. 2018-2024.

Olaszy, G. (2006). *Hangidőtartamok és Időszerkezeti Elemek a Magyar Beszédben* [*Sound Durations and Temporal Structure in Hungarian Speech*]. Akadémiai Kiadó, Budapest.

Ramus, F., M. Nespor & J. Mehler (1999). Correlates of linguistic rhythm in the speech signal. *Cognition* 72, pp. 1-28.

Schiel, F. (1999). Automatic Phonetic Transcription of Non-Prompted Speech. Ohala, J. J. et al. (eds.) *Proc. of the 14th International Congress of Phonetic Sciences*. University of California. San Francisco, pp. 607-610.

Schreiber T. & A. Schmitz (2000). Surrogate time series. *Physica D*. 142, 346.

Tarnóczy, T. (1974). A magánhangzók akusztikai vizsgálatának problémái. [Problems of acoustic analysis of vowels]. *Általános Nyelvészeti Tanulmányok* 10, pp. 153-180.

Váradi, V. & A. Beke (2013). Az artikulációs tempó variabilitása felolvasásban. [The variability of the articulation rate in read speech]. Gósy, M. (ed.) *Beszédkutatás 2013*, pp. 26-41.

White, L. & K. Mády (2008). The long and the short and the final: phonological vowel length and prosodic timing in Hungarian. *Proc. 4th Speech Prosody Conference,* Campinas , pp. 363-366.

Wightman, C.W., S. Shattuck-Hufnagel, M. Ostendorf & P.J. Price (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America* 91, pp. 1707-1717.

Ton(s) d'institutrice. Variation prosodique en invariant situationnel chez une institutrice de maternelle.

Céline Lambeau

celine.lambeau@gmail.com
LEMME - ULg

Abstract

L'article présente une étude exploratoire menée sur des échantillons de parole adressée d'une seule et même institutrice, et destinée à objectiver des différences perçues en écoute non-outillée, d'une part entre le ton qu'elle emploie avec des adultes et celui pris avec les enfants, d'autre part entre les tons employés avec les mêmes enfants dans trois types d'activité différents. L'hypothèse est celle de l'usage par l'institutrice de phonogenres différents dans une même journée. L'analyse, macro-prosodique, porte sur quatre séquences interactives spontanées et indique une correspondance partielle entre le ton de l'institutrice et le phénomène de l'infant-directed-speech, et des variations légères de certains paramètres prosodiques selon le type d'activité menée. L'existence de phonogenres différenciés (selon l'interlocuteur, selon l'activité) n'est donc pas écartée mais devra être confirmée par une étude approfondie d'échantillons de parole plus longs, la durée moyenne de parole articulée pour chaque séquence n'offrant pas une représentativité suffisante pour généraliser les résultats.

1. Entre musical et social

1.1. Cadre de la recherche

Doctorante en Sciences de l'Information et de la Communication, je mène une recherche sur l'interaction musicale, que j'envisage comme un *dispositif de médiation* (Servais 2010) dont l'actant tiers consiste en un matériau musical, susceptible de contribuer par ses caractéristiques immanentes à la définition des relations entre les interactants. Dans ce cadre, j'ai mené une enquête de terrain consacrée aux usages de la musique et du musical par une classe de 1e-2e maternelle. Concrètement, toutes les occurences de son musical (organisé sur le plan de la durée, de la hauteur et du timbre) ont été relevées, et font l'objet d'analyses variées. La prosodie des échanges institutrice-enfants constituant l'un des phénomènes musicaux frappants de ce terrain, une étude prosodique outillée d'enregistrements de la classe est en cours. Cet article rend compte des résultats d'une étude exploratoire portant sur plusieurs échantillons de parole de l'institutrice.

1.2. Infant-directed-speech

En perception globale non outillée, le ton employé par l'institutrice avec les enfants semble relever d'un phénomène prosodique largement documenté depuis 40 ans sous divers labels : *baby-talk* (Ferguson 1964), *motherese/parentese* (Fernald & Kuhl 1987), *parler-bébé* (Imberty 2004), mamanais (Vauclair 2004), ou plus simplement *infant/child-directed-speech* (de Boer 2005). Ce ton adressé aux enfants (IDS), réputé « musical », est largement transculturel et mobilisé de manière intuitive par la plupart des personnes interagissant avec un petit enfant. Il manifeste des différences significatives avec le discours adressé à l'adulte entre autres sur le plan prosodique : fréquence plus élevée, registre vocal plus large, tempo plus lent, rythmicité plus grande (Trainor et al. 2000), contours mélodiques plus marqués (Fernald 1989). On sait aussi que différents types d'IDS sont employés selon le contexte et la nature des activités menées avec les enfants (Trainor et al. 1997). J'ai donc souhaité déterminer si le ton de cette institutrice avec les enfants relevait de l'IDS, et étudier la variation prosodique en rapport avec le type d'activité, afin de mettre au jour un usage éventuel de

phonogenres différenciés (Goldman *et al* 2011).

2. Méthodologie

2.1. Corpus

J'ai réalisé des enregistrements audio de qualité professionnelle de la vie de la classe, au terme d'une immersion longue sur le terrain en qualité d'observateur participant[1]. Quatre séquences discursives impliquant, le même jour, la même institutrice (37 ans, 12 ans de métier) et le même groupe d'enfants (13 filles et 10 garçons, 2,5 à 4 ans) dans des activités différentes en ont été extraites, pour être analysée à des fins exploratoires. Il s'agit d'activités spontanées, habituelles, récurrentes et ritualisées de ce groupe, ce qui permettra une comparaison ultérieure avec les mêmes activités menées par les mêmes acteurs à d'autres dates.

L'une de ces séquences est extraite d'un temps de collation, et montre une alternance d'échanges brefs institutrice-enfant(s) et institutrice-adultes (durée totale : 221 s, dont parole de l'institutrice : 73,8 s – répartition par adresse dans le *Tableau 1*). Toutes les interventions de l'institutrice se manifestant en invariant situationnel[2], cette séquence a été retenue pour son potentiel comparatif quant à l'impact de la dimension de l'adresse dans la variation prosodique (adulte, enfant, groupe).

	Séq. tot.	**ADS**	**IDS**	**GDS**
Durée tot. (s)	221	**28**	**52,2**	9,6
Durée d'art (s)	**73,8**	**18,73**	**44,5**	**7**

Tableau 1 : pour la séquence "collation", répartition des durées de parole et d'articulation de l'institutrice par adresse (adulte, enfant, groupe)

[1] 12 semaines de présence en classe, dans un rôle comparable à celui des stagiaires passifs.

[2] Supervision de la collation, l'institutrice se tenant debout au milieu des tables où mangent les enfants.

Les trois autres séquences correspondent à trois activités discursives menées dans un quartier de la classe dédié aux interactions verbales et musicales (le "coin des amis"), là aussi en invariant situationnel[3]. Ces trois séquences, interactives, ont été labellisées *a priori* comme un "sermon", une "leçon" et une "conversation[4]" selon ma perception globale immédiate sur le terrain. Le temps de parole de l'institutrice représente 40 à 55% de la durée totale de ces séquences (cf.–*Tableau 2*).

Toutes séquences confondues, la durée totale de parole articulée de l'institutrice analysée est de 227 secondes.

	Sermon	**Leçon**	**Convers.**
Durée tot. (s)	**319**	171	191,5
Dur. d'art Instit (s)	**153.9**	**95**	**73,5**
Prop. art. Instit (%)	48,1	55.6	38,4

Tableau 2 : pour les séquences "sermon", "leçon" et "conversation", durée totale des séquences, durée et proportion d'articulation de l'institutrice

2.2. Traitement

Les quatre séquences ont été annotées semi-automatiquement sous Praat (Weeninck & Boersma 2012) selon la méthode exposée par Goldman et al. 2011: transcription orthographique, alignement phonétique par EasyAlign (Goldman 2012a), stylisation de la F0 par Prosogramme (Mertens 2004), annotation manuelle des phénomènes de production, détection automatique des proéminences par ProsoProm (Simon et al 2008), découpage manuel en unités séparées par des pauses (USP). J'y ai ajouté deux tires d'annotation l'une relative au locuteur (institutrice, enfants, adultes), l'autre à

[3] Disposition spatiale en carré, dont un côté accueille l'institutrice, et les trois autres les enfants.

[4] Plus précisément, la séquence se présente comme une succession de micro-conversations institutrice-enfant, au cours desquelles l'enfant est invité à raconter comment s'est passée la fête des pères.

l'adresse, selon que l'institutrice parle à un adulte (ADS), à un enfant (IDS) ou au groupe (GDS).

2.3 Hypothèses

(H1) D'une part, je m'attends à observer plusieurs différences entre ADS et IDS, l'IDS typique manifestant normalement, par rapport à l'ADS, une F0 plus élevée et un débit plus lent. (H2) D'autre part, je fais l'hypothèse de différences objectivables entre sermon, leçon et conversation pour au moins un paramètre prosodique, en postulant en particulier une F0 moyenne sensiblement identique dans les trois séquences, comparable à celle de l'IDS analysée plus haut, un débit plus lent en sermon et leçon qu'en conversation, un taux de parole de l'institutrice plus élevé en leçon et sermon qu'en conversation, et une proportion plus élevée de tons dynamiques montants en conversation.

3. Résultats

3.1. ADS vs IDS

Une table des macro-unités, générée par Prosodyn (Goldman 2012b) à partir de la tire d'adresse des annotations, permet une comparaison des USP sur divers plans. Le *Tableau 3* reprend, par type d'adresse, des mesures relatives à la F0 moyenne, à l'étendue du registre mélodique, et au débit de parole[5]. Une différence de 3,5 ST entre ADS et IDS indique que l'institutrice parle en moyenne une tierce plus haut avec les enfants qu'avec les adultes. La faible différence entre les registres exploités en ADS et en IDS (moins de 2 ST) indique un décalage de la voix vers le haut plutôt qu'un élargissement global de l'étendue vocale.

[5] Mesures effectuées par ProsoReport, voir (Goldman et al, 2008) pour le détail des méthodes de calcul.

	Séq. tot	ADS	IDS	GDS
F0 (ST)	97,5	**95,2**	**98,7**	98
Registre (ST)	10,84	10,17	12	16
Débit (syll/s)	4,7	4,4	5,22	4,3

Tableau 3 : Moyennes de F0, du registre vocal et de la durée des syllabes, pour la totalité de la séquence et par type d'adresse

Les mesures de F0 manifestent donc les caractéristiques attendues de l'IDS. A contrario, on remarque une légère différence quant au débit d'un type d'adresse à l'autre (moins d'une syllabe par seconde), dans le sens d'une rapidité plus grande en IDS, ce qui infirme pour cette séquence l'existence d'un ralentissement net du débit comme autre marqueur de l'IDS.

3.2 Sermon / Leçon / Conversation

Le *Tableau 4* présente les mesures de F0 pour chaque séquence, en regard des mesures de F0 et ADS et en IDS exposées précédemment. Malgré l'adresse "enfant" de ces séquences, les fréquences moyennes durant la leçon et la conversation correspondent plutôt à l'ADS qu'à l'IDS.

	IDS	ADS	Ser.	Leç.	Conv
F0 (ST)	98,7	**95,2**	96,8	**95,9**	**95,4**

Tableau 4 : F0 moyenne et étendue du registre de l'institutrice en IDS, en ADS, durant le sermon, la leçon et la conversation

Devant ce résultat assez inattendu, une réécoute des enregistrements révèle que les enfants auxquels l'institutrice s'adresse durant la séquence extraite du temps de la collation sont les trois plus jeunes de la classe : une moindre autonomie de ces enfants, d'où une tendance de l'institutrice à les materner un peu plus que les autres, pourrait expliquer l'apparition d'un IDS plus typique dans ces interactions que dans les échanges de la séquence conversation.

Quant à la proportion de parole de l'institutrice dans les trois séquences discursives, on peut observer dans le *Tableau 5* qu'elle varie significativement d'une séquence à l'autre, occupant un grand tiers de la conversation et autour de la moitié des séquences leçon et sermon, le reste du temps étant constitué de pauses silencieuses et de parole enfantine en proportions variables, dont l'analyse outillée reste à effectuer. Le débit, par contre, est très comparable d'une séquence à l'autre, ne montrant qu'une très légère accélération en activité conversationnelle.

	Ser.	**Leç.**	**Conv.**
Prop. d'art (%)	48,1	55,6	38,4
Débit (syll/s)	5,1	5	5,6

Tableau 5 : proportion de parole articulée et débit moyen de l'institutrice

La question des contours mélodiques peut être éclairée en partie par des mesures relatives aux tons dynamiques[6]. Le *Tableau 6* montre une variation des proportions et des types de tons dans les différentes séquences.

	Ser.	**Leç.**	**Conv.**
Prop dyn. (%)	12,4	10,4	8,2
Mont. (%)	4.3	**5,3**	**2,4**
Desc. (%)	**8**	5,1	5,8

Tableau 6: proportion de tons dynamiques, montants et descendants

La proportion plus élevée de tons dynamiques dans le sermon pourrait être liée à un facteur émotionnel : dans cette séquence, l'institutrice gronde les enfants, et exprime une contrariété qui semble déborder le cadre strictement didactique de l'exigence du respect des règles. On remarque aussi une présence plus grande des tons descendants, qu'il est difficile d'interpréter sans examen de leur position dans la chaîne linguistique[7].

La proportion de tons dynamiques montants deux fois plus élevée dans la leçon que dans la conversation est inattendue. Ceci s'explique en partie par une tendance de l'institutrice à ponctuer ses interventions par une phrase-mot dont la dernière syllabe est montante et nettement accentuée ("d'accoord ?"). Une étude des courbes intonatives macro-prosodiques pourrait permettre un examen comparatif plus fin des caractéristiques mélodiques des trois séquences.

3.3. Phonogenres ?

Un regard synoptique sur 6 paramètres prosodiques repris dans le ProsoReport (Goldman et al 2008) (*Tableau 7*) indique que chaque paramètre montre des mesures très proches pour deux séquences sur trois (donc varie au moins une fois) et que chaque séquence manifeste une différence avec les deux autres pour au moins un paramètre. Aucune des mesures ne montrant une coïncidence complète à travers les trois séquences, la possibilité de se trouver face à trois phonogenres différenciés n'est pas écartée. Des comparaisons avec d'autres séquences du corpus complet spontanément labellisées comme sermons, leçons ou conversations devront par conséquent être effectuées pour confirmer cette possibilité.

	Ser.	**Leç.**	**Conv.**
F0 (ST)	**96,8**	95,9	95,4
Débit (syll/s)	5,1	5	**5,6**
Registre (ST)	12,5	**9,2**	12
Tons dyn (%).	12,4	10,4	**8,2**
Proém (%).	**29**	24,2	24,3

[6] Tels que calculés par ProsoReport, (Goldman et al 2008).

[7] Un tel examen nécessite le recours à une annotation grammaticale effectuée selon la méthode décrite par (Goldman *et al* 2009), non réalisée à ce jour pour les séquences étudiées ici.

Agitation (ST/s)	9,1	9,2	**8,6**

Tableau 7: moyenne de la F0, du débit, du registre vocal, des proportions de tons dynamiques et de proéminences et de l'agitation mélodique

4. Conclusion

Au terme de cette analyse exploratoire, l'hypothèse (H1) d'une différence entre IDS et ADS n'est que partiellement confirmée, les mesures ne montrant une F0 nettement plus élevée que dans la parole adressée aux enfants les plus jeunes durant la collation, tandis que le débit est légèrement plus rapide avec les enfants qu'avec les adultes contrairement aux prédictions basées sur les traits connus de l'IDS. Une confrontation ultérieure avec les caractéristiques connues du *teachertalk* plutôt que du *babytalk* est dès lors à envisager. L'hypothèse (H2) de différences objectivables entre les trois séquences pour au moins un paramètre prosodique est globalement confirmée, avec des variations inverses de celles attendues pour deux des quatre sous-hypothèses. L'existence d'idiogenres différenciés n'étant pas écartée, l'étude exploratoire confirme l'intérêt d'une étude approfondie de la variation prosodique chez cette locutrice. Celle-ci devra cependant mobiliser des échantillons de parole plus longs pour chaque phonogenre postulé, pour atteindre aux conditions permettant une représentativité.

Remerciements

Je tiens à remercier Mme Anne-Catherine Simon, pour la formation reçue en matière d'analyse prosodique outillée, et pour ses conseils et sa disponibilité aux différentes étapes de traitement de mes données.

Références

de Boer, B. (2005) Infant-Directed Speech and Evolution of Language. In : Maggie Tallerman (Ed.), *Language Origins: Perspectives on Evolution*. Oxford University Press.

Ferguson, C. (1964), Baby talk in six languages. *American Anthropologist*, 66, pp. 103–114.

Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants : Is the melody the message ? *Child Development*, 60, pp. 1497-1510

Fernald, A. & Kuhl, P. (1987). Acoustic determinants of infant preference for motherese speech. *Infant Behavior and Development*, 10, pp. 279-293.

Imberty, M., (2004), Le bébé et le musical. In : Nattiez, J-J (Ed), *Musiques. Une encyclopédie pour le XXIe siècle,* Actes Sud/cité de la Musique, Paris.

Goldman J.-P. (2012a), EasyAlign [Computer program], http://latlcui.unige.ch/phonetique.

Goldman, J.-P. (2012b). ProsoDyn: a graphical representation of macroprosody for phonostylistic ambiance change detection. *Proceedings of the 6th International Conference on Speech Prosody 2012* (p. 75-78). Shangaï (China): Tongji University Press.

Goldman, J-P, Auchlin, A., Simon, A.C., (2011). Description prosodique semi-automatique et discrimination de styles de parole. In: Yoo, H-Y & Delais-Roussarie, E. (eds), *Actes d'IDP 2009*. Paris, pp. 207-221

Goldman, J.-P., Auchlin, A. Avanzi, M. & Simon, A. C., (2008) ProsoReport: an automatic tool for prosodic description. Application to a radio style, *Proceedings of Speech Prosody'08*. pp. 701-704

Mertens P. (2004), Le prosogramme : une transcription semi-automatique de la prosodie, *Cahiers del'Institut de Linguistique de Louvain,* 30 : 1-3, pp. 7-25.

Servais C., (2010), Qui dispose des dispositifs ?, *Questions de communication*, 10, pp. 7-16

Simon, A.C., Avanzi, M. & Goldman, J.-P. (2008), La détection des proéminences syllabiques. Un aller-retour entre l'annotation manuelle et le traitement automatique », *Actes du 1er Congrès Mondial de Linguistique Française),*. pp. 1673-1686

Trainor, L.J., Clark, E.D., Huntley, A., & Adams, B. (1997). The acoustic basis of infant preferences for infant-directed singing. *Infant Behavior and Development,* 20, pp. 383-396

Trainor L-J, Austin C.M., & Desjardins R. N. (2000), Is infant-directed-speech the result of the expression of emotions ?, *Psychological Science* 11 : 3, pp. 188-195

Vauclair, J. (2004), *Développement du jeune enfant, motricité, perception, cognition*. Paris, Belin.

Weenink, D. et Boersma, P., (2012). Praat: doing phonetics by computer (Version 5.3.23) [Computer program], http://www.praat.org

Routes to Prominence in Free Word Order Language Discourse

Tatiana Luchkina[1] *& Jennifer S. Cole*[2]

luchkin1@illinois.edu, jscole@illinois.edu
[1, 2]University of Illinois at Urbana-Champaign

Abstract
Perceived prominence in Russian, a free word order language, can be communicated prosodically and/or via word order. Paired production and perception experiments with native speakers show that discourse-prominent constituents are marked acoustically and through a change in word order. These two route to prominence may reinforce each other, as evident from distinctively higher duration and intensity values associated with ex-situ words, as well as their higher visibility in discourse.

1. Introduction

An essential aspect of comprehending language, in written or spoken modalities, is interpreting the status of a linguistic entity relative to the discourse or narrative context. A word or phrase can be introduced as information that is new to the discourse and as relatively important or emphasized, or as given information of lesser significance, and languages may express discourse status by means of prosody, morphology, and the sequencing of constituents within the sentence [1]. This paper investigates the relationship between phrasal prosody and word order in the expression of information structure (IS) in Russian, a free word order language. Results from speech production experiments, and speech and text comprehension experiments are examined to test whether prosody and word order are typically used independently or together in the encoding and decoding of IS in discourse.

Russian is chosen as the test case because first, it allows but does not require surface reordering of sentential constituents for information structural purposes, and second, it exhibits distinctions in prosodic prominence among the constituents of a sentence [2,3].

1.1. IS and Prominence

In languages like English and German, it is generally the case that a word can be assigned prosodic prominence as an expression of its discourse status regardless of its position within the utterance, i.e., *in situ* [4,5]. While the phrase-final position is the default location of the primary prominence in English, the phenomenon of metrical reversal shown in (1-2) illustrates prominence displacement, where the primary prominence shifts leftwards to signal the status of the prominent entity as new or with contrastive focus (examples from [6].

(1) Joel bought a green CAR.[1]
(2) Joel bought a GREEN car.

Apart from prosodic marking, the sequencing of constituents within the utterance or sentence presents an alternative route for encoding IS in so-called free word order languages, eg. in the Italian sentences in (3) below (from [7]), by dislocating the subject noun from its canonical (pre-verbal) position, a speaker or writer deploys the tool of *ex-situ prominence* [7,8] thereby marking the subject as focused and prominent:

(3) E' arrivato MARIO.
arrived Mario

Given that prosody and word order provide two means for encoding IS, it follows that some languages may utilize both mechanisms. This prediction is confirmed for Georgian [8], where the distribution of in- and ex-situ prominence maps onto two different foci types: Ex-situ contrastive or exhaustive focus and in-situ prosodically prominent informational focus. Dual

[1] Here and below prominent words appear in capital letters.

marking of discourse prominence has been reported for the variety of Romani spoken in Komotini, where focus marking is accomplished via a combination of prosodic and morphosyntactic strategies employed in the same utterance [9].

1.2. Encoding IS in a free word order language

For languages that may utilize more than one strategy to mark discourse-prominent information, we may ask how structural and/or prosodic means may apply selectively or concurrently in order to communicate different categories of information in discourse. Here we explore this issue for Russian. The semantically neutral, default word order in Russian is SVO, and as in other free word order languages, a word can appear in its canonical position (in-situ), fronted, or post-posed. The ordering of the constituents in a sentence marks IS and not grammatical function. To illustrate, both (a) and (b) are possible continuations for the sentence in (5), in different discourse conditions. In the context provided in (5), the word *Ivan*, critical to the understanding of who is doing the cooking, may be located in the rightmost position, where it is structurally prominent (as in b), or it may occur pre-verbally as in (a).

(5) Tri druga, Ivan, Petr, i Andrey, nahsli novjiy retsept pizzj.
Three friends, Ivan Petr and Andrey, found a new pizza recipe.

a. IVAN gotovit pizzu.
Ivan-SUBJ cooks pizza-OBJ

b. Pizzu gotovit Ivan.
pizza-OBJ cooks Ivan-SUBJ

In this study we test the hypothesis that the difference in word order illustrated in (5) is accompanied by a difference in prosodic prominence (assigned to *Ivan* in this example). We explore how prosody and word order function independently and in combination to mark IS in Russian.

A production experiment and an off-line perception experiment were administered to determine (1) whether the perceived prominence of a word is dependent on its sentential position; (2) whether non-canonical word order by itself is a sufficient tool to mark IS, (3) whether concurrent prosodic marking of ex-situ constituents is necessary to confer greater discourse salience.

Our general hypothesis is that positioning of a word in a designated ex-situ position is an independent cue to prominence, which may be further reinforced with acoustic-prosodic features associated with such position. To test this hypothesis, a miniature corpus comprised of two authentic Russian texts is analyzed for word order and IS properties. Read productions of these texts are analyzed for acoustic evidence of prosodic marking in relation to word order and IS. Perceptual rating studies were performed using written and spoken utterances. Experimental results are analyzed for the relationship between word order, prosodic marking and perceived prominence.

2. **Experiment 1: Production task**

2.1. Method

To obtain a range of word order and prosodic discourse features, two published narratives were read orally by 8 female speakers of Russian (ages 21-38). The miniature corpus included a short folk tale (text A) and an excerpt from the biography of a Russian poet (text B). With an average sentence length of 5.2 content words (SD =1.77), approximately 30% of the sentences in the chosen narratives deviate from the canonical SVO order.

Following a discourse annotation framework introduced in [10], the miniature corpus was annotated for four kinds of IS categories: ‘THEME’ (discourse-given), ‘RHEME’ (discourse-new), ‘MEDIATOR’ (inferable), and ‘CONNECTOR’ (function words). Annotations were independently done by one of the authors (TL) and a second native

speaker of Russian. Inter-rater agreement (linearly weighted Kappa) between the annotators, across texts was very strong: $\varkappa$=0.89, SE=0.03, α=0.05. [2] Words were also marked for Focus, and for sentential position as in-situ or ex-situ (specifically, 'Fronted' or 'Post-posed', relative to SVO order).

The observed distribution of IS categories in different sentence positions differed by text: while a reliable association held between the features 'fronted' (an ex-situ word is positioned sentence-initially) and 'RHEME', no such association held between features 'Fronted' and 'THEME', Pearson χ^2=11.26, P=0.001). Such association between the surface order of the constituents and their IS category provides evidence that word order variability in the Russian corpus is used to promote novel information to the clause-initial position.

Acoustic measures were examined as correlates of prosodic prominence, and analyzed for their relationship to the IS and sentential position of a word. The acoustic-prosodic measures of f0 (Hz) and intensity maxima, and vowel duration were taken from the stressed syllable of each IS-coded content word in the corpus for a total of 230 words.[3] The relationship between normalized averaged acoustic measures, IS category, and sentential position of a word was then tested in a series of linear, mixed effects multivariate regression analyses.

2.2. Results

Successful predictors of the acoustic measures extracted from the corpus (see Table1) include RHEME, which is associated with higher mean values of intensity and duration, and Focus, associated with significantly higher f0 maxima. All acoustic measures, i.e., max intensity, duration, and max f0 are successfully predicted by the sentential position of the target word, with significantly higher values for words that appear ex-situ and fronted.

Table 1: *Predictors of the stressed vowel intensity, duration, and f0, with respect to the carrier word*[4]*:*

(max)intensity	(max)f0	vowel duration
ex-situ, fronted (t=2.01, p<0.05)	ex-situ, fronted (t=2.13, p=0.03)	ex-situ (t=2.01, p=0.05)
RHEME (t=1.93, p=0.055),	focus(emphatic/ contrastive) (t=2.68, p=0.01)	RHEME (t=2.94, p<0.01)
*CONNECTOR (t=-3.1, p<0.01)	*MEDIATOR (t=-2.72, p=0.01)	*CONNECTOR (t=-3.08, p<0.01)

*This factor predicts a lower value for the parameter of interest.

3. Experiment 2: Prominence Rating Task

Analysis of the production data established that the acoustic parameters of f0, intensity, and duration vary in relation to the IS category of a word and focus, and are also associated with its sentential position in the sentence. To determine if some/all of these parameters significantly affect listener's perception of a word as prominent, i.e., as important in relation to the discourse meaning, structural and acoustic-prosodic cues to prominence were determined on the basis of reading and auditory comprehension tasks performed by linguistically naïve native speakers of Russian (N=49 (reading modality), N=27 (auditory modality)).

3.1. Method

An offline perception task was conducted with 39 clause-size excerpts from the

[2] Maximum possible $\varkappa$, given observed marginal frequency=0.94.

[3] The values of max f0 and max intensity were taken from the center region of the vowel, excluding 20 ms from the left and right edges of the vowel as identified by acoustic criteria, and normalized within-utterance.

[4] Factors '*speaker*' and '*word*' (not shown in Table1) were included in the model as random effects.

miniature corpus. A clause (vs. an intonation phrase) was chosen as a unit of presentation as it expresses one relatively complete idea and can be perceived as a whole. Each clause, or target segment, was presented along with the preceding context. The mode of presentation was either written text or audio recording performed by one female speaker of Russian (age=28). Respondents read the entire portion of the text preceding the target segment, read or listened to the target segment and identified discourse-prominent word(s) in the target segment by associating them with one level of the binary variable "+/- prominent". Following [11], no formal definition of prominence was given. Participants were instructed to mark only those words that 'were the focus of their attention' in the utterance, based on the preceding context. Any number of content words could be marked as prominent.

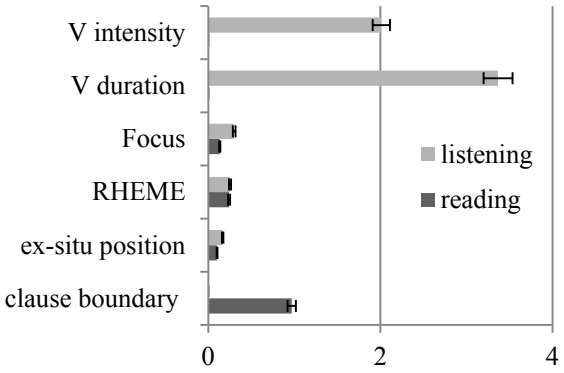
3.2. Results

Consistency of the responses: Responses to the prominence rating task were assessed for inter-rater agreement. The agreement coefficients obtained translate into fair, though highly significant agreement levels: Fleiss' kappa=0.26 ($p<0.001$) for the written and 0.36 ($p<0.001$) for the auditory modality.

Correlates of perceived prominence: Following [12], each word in the narratives was assigned two discourse salience scores (one per test modality), which were obtained by dividing the total number of times a word was chosen as salient by the total number of participants who responded to the relevant test question. These prominence scores were used as a quasi-continuous measure of perceived prominence, and were submitted to linear regression analyses in which predictors for prominence perception included those listed in Table 1, with the variables *'participant'* and *'test item'* as random effects.

In silent reading of the Russian corpus, words associated with higher prominence scores were those that were located at clausal boundaries (t=1.97, p=0.05), in ex-situ position (t=2.08, $p<0.05$), carrying new information (t=4.55, $p<0.001$), and focused contrastively or emphatically (t=2.61, p=0.01). In the auditory modality, these factors are complemented with two acoustic predictors, duration (t=2.02, p=0.05) and intensity (t=2.24, $p<0.05$). Figure 1 summarizes the findings of the perceived prominence analyses.

Figure 1: Weight of significant predictors of perceived prominence by modality: *Predictor weights (x-axis) as determined by the associated regression coefficient for the significant predictors from the rating task (y-axis).*



4. Discussion

Analysis of perceived prominence in Russian was conducted to determine which factors guide naïve readers' or listeners' perception of a word as prominent in a discourse or narrative. A special point of interest was to determine whether variation in word order can be utilized as a means of encoding the information status of a word and its perceived prominence. Results demonstrate that independent of the modality of presentation, words associated with new information or Focus are perceived as more prominent. In a free word order language such as Russian, information status is encoded via two routes, prosodic and syntactic.

In the auditory modality, listeners treat the acoustic-prosodic realization of a word as a cue to its discourse status. This is evident from the finding that greater duration and intensity of the stressed vowel reliably trigger perception of a word as prominent. About 30% of the utterances in the mini corpus of published narratives used in this work deviate from canonical word order. Results of the prominence rating task show that apart from the acoustic effects of prosody, an ex-situ position of a word also contributes to its perception as prominent. Analysis of the syntactic and acoustic-prosodic characteristics of perceived prominence reveals that different cues to prominence may apply concurrently: When the word is situated non-canonically, it is more likely to have a higher prominence score and, in the auditory modality, to perceptually stand out by virtue of having greater duration and intensity. The distinctive acoustic realization of a non-canonically positioned content word may not only cue its relatively high informational load and discourse prominence, but may also (redundantly) signal that the word is left- or right- dislocated. This is evident from the finding that most of the prominence predictors are selectively associated with clause-initial and clause-final sentential positions which may be reserved for words carrying higher informational load. This result is consistent with the previous findings that Russian exhibits focus fronting and right-edge dislocation for IS purposes [3,13].

5. Conclusion

This study contributes to the understanding of discourse-prominence in a free word order language. While results of the production and perception experiments performed by linguistically naïve native speakers of Russian reveal that perceptually salient acoustic-prosodic realization of discourse-prominent information holds under free word order, further studies are necessary to determine whether cross-application of prominence cues is characteristic of all vs. select categories of discourse-prominent information and whether its effect is additive, i.e., leading to a word being associated with a yet greater degree of perceived prominence.

References

[1] Morgan, J. L., Meier, R. P., & Newport, E. L. (1987). Structural packaging in the input to language learning: Contributions of prosodic and morphological marking of phrases to the acquisition of language. Cognitive Psychology, 19, 498–550.

[2] Sekerina, I. (2003). Scrambling processing: Dependencies, complexity, and constraints. In S. Karimi (ed.), Word order and scrambling UK: Blackwell, 301–324.

[3] Slioussar, N. (2011a). Processing of a free word order language: The Role of Syntax and Context. Journal of Psycholinguistic Research, 40:291-306.

[4] Ladd, R. D. (2008). Intonational Phonology. Cambridge University Press.

[5] Katz, J., Selkirk, E. (2011). Contrastive focus vs. discourse-new: Evidence from phonetic prominence in English. Language, 87, 771-816.

[6] Calhoun, S. (2010). The Centrality of Metrical Structure in Signaling Information Structure: A Probabilistic Perspective, *Language*, 86(1), 1-42.

[7] Donati, C. & N. Nespor (2003). From Focus to Syntax. Lingua, 113-11, 1119-42.

[8] Skopeteas, S. & Fanselow, G. (2010). Focus in Georgian and the expression of contrast. Lingua, 120, 1370-1391.

[9] Arvaniti, A. & Adamou, E. (2011). Focus expression in Romani. In Proceedings of the 28th West Coast conference on Formal Linguistics, Somerville, MA: Cascadilla Proceedings Project.

[10] Calhoun, S., M. Nissim, M. Steedman, and J.M. Brenier. (2005). A framework for annotating information structure in discourse: Pie in the Sky. Proceedings of the workshop, ACL, 45–52.

[11] Cole, J., Mo, Y., & Hasegawa-Johnson, M. (2011). Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology*, v.1, 2011, p. 425-452.

[12] Mo, Y., Cole, J., & Lee, E. (2008). Native listeners? Prominence and boundary perception. Speech Prosody 2008.

[13] Neeleman, A., Titov, E. (2009). Focus, contrast, and stress in Russian. *Linguistic Inquiry* 40, 514–524.

Phrase-initial boundary tones in Hungarian interrogatives and exclamatives

Katalin Mády, Beáta Gyuris, and Ádám Szalontai

Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest
{mady|gyuris|szalontai}@nytud.hu

Abstract

There is a group of *wh*-interrogatives and *wh*-exclamatives in Hungarian that are distinguished only by means of prosody. It was shown previously that the distinction consists in having falling pitch accents on the *wh*-element in interrogatives, and rising pitch accents in exclamatives. In this paper, the relevance of sentence-initial f0 is investigated as a potential trigger for the above differences. A perception experiment was set up in which sentence-initial and sentence-final chunks containing only f0 information were presented to participants, along with the *wh*-element bearing the only pitch accent of the sentence. It was shown that sentence-initial chunks carried the most relevant information for sentence type identification, whereas pitch accent type and sentence-final f0 were less informative. The findings suggest that phrase-initial boundary tones are of relevance in Hungarian prosody.

1. Introduction

Hungarian prosody is left-headed: lexical stress is fixed to the word-initial syllable, while pitch accents initiate a lower level phrase (Varga 2002, Hunyadi 2002). It is not clear whether there is a default position for nuclear accents in the Hungarian sentence. (É. Kiss 2002, p. 11) claims that "[i]n Hungarian, phrasal stress – similar to word stress – falls on the left edge, i.e., the Nuclear Stress Rule of Chomsky & Halle (1968) operates in a direction opposite to that attested in English." One manifestation of this phenomenon seems to be that the most prominent unit is the left-most element of the second (obligatory) major part of the sentence, the predicate part. (The first major part, the topic, can also be missing.) The predicate part includes the syntactically expressed focus, which is situated in a position immediately preceding the verb. The presence of a focus constituent forces deaccentuation of the verb and of the following postverbal elements within the same prosodic unit.

Hungarian does not make extensive use of phrase-final boundary tones: H% tones occur only as a continuation rise, but they are not utilised for marking question intonation. However, there are certain sentence types and/or illocutionary forces that are distinguished purely by means of intonation. One example is the default type of yes/no interrogatives that are string-identical with declaratives. Prosodically, they are characterised by an underlying L* H L% contour (Ladd 2008, p. 182), whereas the penultimate H is missing in declaratives. Another example for a purely prosodic distinction is the case of *wh*-interrogatives and a particular type of *wh*-exclamatives, which will be discussed in detail below.

The goal of the present paper is to clarify which prosodic units contribute to the distinction between root wh-interrogatives and wh-exclamatives. First, a general outline of the syntactic structure of these sentence types is given. Then the results of a recent production experiment are presented. Finally, we further test our hypotheses from a previous experiment with new perceptual data.

1.1. The syntax of root *wh*-interrogatives and wh-exclamatives in Hungarian

The *wh*-expressions in *wh*-interrogatives in Hungarian, illustrated in (1), are standardly assumed to occupy the syntactic focus position of the sentence, considered to be a specifier of a Focus Phrase (FocP) within the hierarchically structured preverbal field, shown in (2), a simplified version of Lipták (2006, p. 362, ex. (40)). (Cf. É. Kiss 2002 for discussion.)

If the latter position is filled, the verb moves to the head of FocP, to be adjacent to the focus, leaving the verbal prefix referred to as 'pv' behind, which is situated immediately in front of the verb in neutral sentences.) Thus, in what follows, the verb-prefix order will be referred to as one involving *inversion*.

(1) Mennyire éhezett meg János?
how.much grew.hungry.3sg pv János
'How hungry did János get?'

(2) $\ldots [_{FocP}\ \{\text{focus}\}\ [_{Foc'}\ V^0\ [_{AspP}\ \text{pv}\ \ldots]]]$

The current paper is concerned with one of the three types of root exclamatives distinguished in Lipták (2006), the so-called *wh*-exclamatives. Lipták (2006) classifies *wh*-expressions into three groups, depending on how the structure of the exclamatives they appear in relates to that of the corresponding *wh*-interrogatives.

The first group of *wh*-expressions require the verb and the prefix to occur in the non-inverted order, as in (3), making it necessarily different from the corresponding interrogative, in (1):

(3) Mennyire megéhezett János!
how.much pv.grew.hungry.3sg János
'How hungry János became!'

Lipták (2006) argues that in examples like (3) the *wh*-expression (Exclamative Phrase) is not in Spec,FocP but in a position immediately dominating the latter.

The second group of *wh*-expressions gives rise to necessarily string-identical *wh*-exclamatives and interrogatives, as (4)-(6) show:

(4) Milyen későn kelt fel?
how late got.up.3sg pv
'How late was it when he got up?'

(5) Milyen későn kelt fel!
how late got.up.3sg pv
'How late it was when he got up!'

(6) *Milyen későn felkelt!
how late pv.got.up.3sg
Intended: 'How late it was when he got up!'

Lipták (2006) assumes that in sentences like (5) the *wh*-expression also occupies the focus position.

A third group of *wh*-expressions gives rise to grammatical exclamatives both with and without verb-prefix inversion, illustrated in (7)-(8).

(7) Hány almát ettél meg!
how.many apple.acc ate.2sg pv
'You ate so many apples!'

(8) Hány almát megettél!
how.many apple.acc pv.ate.2sg
'You ate so many apples!'

Lipták (2006) considers both form types illustrated in (7) and (8) as representatives of the *exclamative* sentence-type, which have an identical prosodic form, consisting of a "stress on the E[xclamative]-phrase and falling intonation following it" (p. 345, fn. 3). In Kálmán (2001, p. 137) the prosody of *wh*-exclamatives is characterized as a "high tone followed by a slow descent".

1.2. The prosody of *wh*-interrogatives and *wh*-exclamatives in Hungarian

In order to test the accuracy of the above claims on prosody, the three types of *wh*-exclamatives were investigated in a production experiment by Gyuris & Mády (2013) recently. All sentences started with a *wh*-expression and included inverted word order for interrogatives, and inverted, non-inverted word order or both for exclamatives. The goal was to investigate whether *wh*-interrogatives and *wh*-exclamatives are distinguished (1) in terms of tonal categories, (2) in terms of absolute f0 values, and to see whether (3) differences between the three types of exclamatives lead to differences in the distinction.

In terms of tonal labels, *wh*-interrogatives typically started with a high tone followed by a high or a falling pitch accent and a low phrase-final boundary tone, whereas *wh*-exclamatives started with a mid or low initial tone followed by a rising pitch accent and a mid final boundary tone. The categorical labels were only partly reflected by the parametric analysis: both the sentence-initial f0 and the f0 maximum were significantly higher in interrogatives than in exclamatives, but sentence-final f0 did not differ between the two sentence types. This suggests that the perception of a mid tone in exclamatives is a result of a lower phrase-initial f0 or the lower f0 maximum within the sentence. None of the categories investigated showed any difference between the three syntactic subtypes of *wh*-exclamatives (involving inversion, no inversion or optional inversion, respectively).

This experiment shows that *wh*-interrogatives and *wh*-exclamatives mainly differed with respect to their pitch accent patterns. *Wh*-interrogatives were

previously found to bear a falling accent by Mycock (2010). She also claimed that the *wh*-word can optionally be preceded by a high tone (p. 284).

A wide range of experiments on several languages have shown that several prosodic entities can be utilised for distinguishing between sentence types and/or speech acts, such as nuclear or prenuclear accents or boundary tones. The revised version of Sp-ToBI links L+H* nuclear accents to exclamatives (Estebas Vilaplana & Prieto 2009). Prenuclear accents were found to be relevant in Neapolitan Italian where the initial part of a sentence led to a reliable distinction between yes/no questions and statements (Petrone & D'Imperio 2011). In other varieties of Italian, higher sentence-initial f0 was found to accompany non-*wh* exclamatives when compared to broad focus declaratives. Other studies concluded that sentence types is expressed by the interplay of several prosodic factors (Batliner 1989).

The above studies show that languages use different prosodic means for expressing sentence type and/or illocutionary force, and most of them agree that some prosodic units play a more important part than others. In the perception experiment to be presented here, we investigated the question of how the differences between Hungarian *wh*-interrogatives and *wh*-exclamatives can be modelled. The study had the following goals: Can *wh*-interrogatives and *wh*-exclamatives be distinguished by their (1) sentence-initial f0, (2) the pattern of the pitch accent, or (3) their phrase-final f0? Furthermore, (4) the effect of identical vs. different word orders on the identification accuracy were tested.

2. Materials and methods

There were eleven pairs of target sentences, each pair consisting of an interrogative and a root exclamative. The structure of interrogatives followed the following pattern, where DM_1 and DM_2 refer to unaccented discourse markers having a total length of 4 syllables:

(9) [DM_1 DM_2 Wh-expression V pv]

The *wh*-expression either consisted of a single *wh*-word or a *wh*-word+adjective/noun phrase.

The structure of root exclamatives followed two patterns. Those containing *wh*-expressions only compatible with the inverted word order followed the pattern shown in (9) (5 examples). Those with *wh*-expressions compatible with both orders followed the pattern in (9) in 3 cases, and followed the pattern shown in (10) in the remaining 3 cases:

(10) [DM_1 DM_2 Wh-expression pv V]

(11)-(12) illustrates a pair with inverted word order. Capitals indicate pitch accent.

(11) Na de akkor MIlyen későn kelt fel?
so but then how late got.up.3sg pv
'But then how late was it when he got up?'

(12) De hogy végül MIlyen későn kelt fel!
but that finally how late got.3sg up
'But eventually how late it was by the time he got up!'

Target sentences were spoken by a male speaker. Since it is not possible to use identical particles for interrogatives and exclamatives, sentences were transformed so that segmental and intensity cues can be eliminated. Sound samples were edited in Praat 5.3.40: first, f0 movements due to microprosodic changes (e.g. higher f0 onsets after unvoiced consonants) were corrected manually. Subsequently, the entire sentence was synthesised into a so-called "hum", a human-like schwa-sound. Three segments were cut from these sound files: the initial 3 syllables of the discourse markers that contained no substantial f0 movement (*ini*), the *wh*-element (*med*) and the final 2 syllables that again had a relatively flat f0 curve (*fin*). The f0 values of the stimulus sentences were not changed, thus they reflected the speaker's original production of the sentences.

The recorded sounds of the two different types of sentences were compared in terms of their initial, final and maximal f0 values. Table 1 shows the range and the mean of the relevant measures where range is the difference in f0 between the minimum and the maximum in the *med* segment, the position of the pitch accent. Figures 1 and 2 show typical f0 contours of the hummed samples.

Table 1: Mean f0 in original stimuli (Hz)

	initial	final	max.
excl	125	117	176
int	150	130	213
$p <$	0.001	0.05	0.001

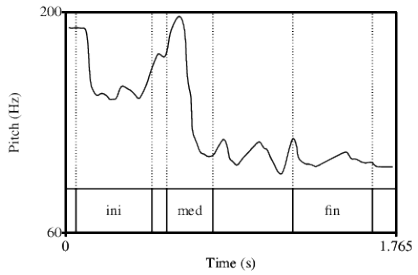


Figure 1: F0 contour of the hummed sample of an interrogative.

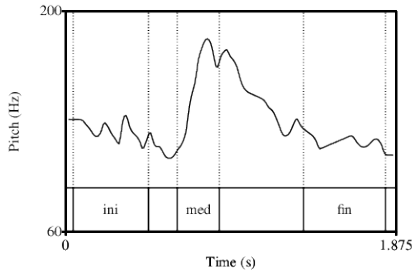


Figure 2: F0 contour of the hummed sample of an exclamative.

The results coincide with the findings of the previous experiment presented above: interrogatives had significantly higher f0 values all throughout the sentence but especially sentence-initially. The position of the maximal peak within the *med* segment was also investigated. On a scale where 0 marks the beginning of the segment and 1 its end, the mean value of the position of the peak was 0.77 in exclamatives and 0.33 in interrogatives ($p = 0.001182$). These results also parallel the production experiment outlined above where *wh*-interrogatives had falling while *wh*-exclamatives had rising pitch accents. In exclamatives, the overall f0 remained higher between the accented syllable and the end of the stimulus, which corresponds to the presence of rising pitch accents in this sentence type.

Participants were presented with three chunks from each sentence containing the initial part that represented the phrase-initial boundary tone, the *wh*-element bearing the pitch accent, and the final part representing the phrase-final boundary tone. While presented with the chunk, they saw two entire sentences on the screen, distinguished both by the initial particles and the appropriate punctuation mark. The position of the chunk currently heard was indicated by an arrow below the corresponding part of the sentences on the screen for both sentences. Participants had to decide in a binary forced choice task whether they were listening to a chunk from an interrogative or an exclamative sentence. Additionally, 22 filler sentences were included. In order to reduce the monotony of the task, in some filler samples participants heard the original speech sample and not the synthesised hum. Target sentences were presented in an individualised random order, preceded by a training phase. There was a total of 24 subjects (7 females, 17 males, mean age 42 years).

3. Results

The analysis is based on the distribution of correctly identified utterances over sentence types and the position of the chunks. Since the sentence-initial chunk includes unaccented syllables only, it correlates with a phrase-initial boundary tone. The sentence-medial chunks were always identical with the *wh*-expressions (one or two syllables). The sentence-final chunks again included unaccented syllables with no or little pitch movement within the sequence, thus they were correlated with a phrase-final boundary tone. Differences between sentence types, i. e. the homogeneities of the distributions between them were tested by means of χ^2 tests. Differences between the number of correct identifications for each subject were compared by repeated measures ANOVA. The significance level was set to $p < 0.05$.

As shown in Figure 3, the number of correctly identified chunks was unevenly distributed both among sentence types and sentence positions. The distributions of correctly identified interrogatives vs. exclamatives over chunk positions were inhomogeneous according to a χ^2 test ($p < 0.0001$). The initial chunk is the only one that yields correct identifications for both sentence types above chance level (50%). The results show indirectly that there was a strong bias towards exclamatives when final chunks were presented, which points to an overall uncertainty regarding this position. The lower identification rate for medial segments might be due to the fact that the f0 peak is often delayed and is located behind the pitch-accented *wh*-element.

The impact of chunk position on the number of correct identifications was investigated for each subject separately, by means of repeated measures

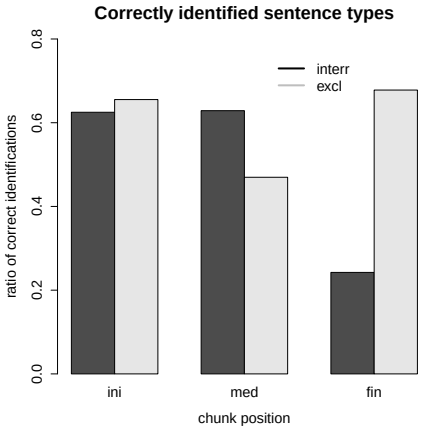


Figure 3: Distribution of correct responses for interrogatives and exclamatives depending on initial, medial and final chunk position.

ANOVAs, the number of correct identifications as the dependent variable, and position as a within-subject factor. The analysis showed a highly significant difference for interrogatives ($p < 0.0001$), and for exclamatives ($p = 0.003$), where the number of correctly identified final chunks is the lowest for both sentence types.

4. Discussion

The above results demonstrate that sentence-initial f0, the f0 pattern of the (only) pitch accent, and sentence-final f0 contribute to the identification of interrogatives and exclamatives to a different extent. The most reliable measure for the distinction was sentence-initial f0 that triggered correct identifications in over 60% of all cases. This shows that the unaccented part of these sentences carries more relevant information with regard to the sentence type than the pitch accent pattern or sentence-final f0.

These findings can be interpreted as a hint to the presence of phrase-initial boundary tones in Hungarian: a %H tone for interrogatives, as was already suggested by Mycock (2010), and a %L tone for exclamatives. The relevance of phrase-initial boundary tones is in line with the fact that important information within the Hungarian sentence is typically located close to the left edge of the sentence. In other words, left-headedness in syntax and prosody seem to enhance the concentration of informational weight towards the left edge of higher syntactic and prosodic units.

5. Acknowledgements

This research was supported by the Hungarian Scientific Research Fund (PD 101050 and NK 100804) and the Momentum program of the Hungarian Academy of Sciences *Division of labour between linguistic subsystems in the expression of quantifier scope*.

6. References

Batliner, A. (1989), Wieviel Halbtöne braucht die Frage? merkmale, Dimensionen, Kategorien., *in* H. Altmann, A. Batliner & W. Oppenrieder, eds, 'Zur Intonation von Modus und Fokus im Deutschen', Niemeyer, Tübingen, pp. 111–162. [How many semitones does a question need? Features, dimensions, categories].

Chomsky, N. & Halle, M. (1968), *The Sound Pattern of English*, Harper and Row, New York.

É. Kiss, K. (2002), *Hungarian Syntax*, Cambridge University Press, Cambridge.

Estebas Vilaplana, E. & Prieto, P. (2009), 'La notación prosódica en español. una revisión del Sp_ToBI', *Estudios de Fonética Experimental* **XVIII**, 263–283.

Gyuris, B. & Mády, K. (2013), Approaching the prosody of Hungarian wh-exclamatives, *in* P. Szigetvári, ed., 'VLLXX: Papers presenpted to László Varga on his 70th birthday'. http://seas3.elte.hu/tmp/vlfs/gyurismady.html.

Hunyadi, L. (2002), *Hungarian sentence prosody and universal grammar: on the phonology–syntax interface*, Lang, Frankfurt/Main.

Kálmán, L. (2001), *Magyar leíró nyelvtan 1. Mondattan*, Tinta Könyvkiadó, Budapest. [Hungarian descriptive grammar 1. Syntax].

Ladd, D. R. (2008), *Intonational phonology, 2nd ed.*, Cambridge University Press, Cambridge.

Lipták, A. (2006), 'Word order in Hungarian exclamatives', *Acta Linguistica Hungarica* **53**, 343–391.

Mycock, L. (2010), 'Prominence in Hungarian: the prosody–syntax connection', *Transactions of the Philological Society* **108**(3), 265–297.

Petrone, C. & D'Imperio, M. (2011), From tones to tunes: Effects of the f0 prenuclear region in the perception of Neapolitan statements and questions, *in* 'Prosodic categories: production, perception and comprehension', Springer, New York, pp. 207–230.

Varga, L. (2002), *Intonation and stress: evidence from Hungarian*, Palgrave Macmillan, Basingstoke & New York.

Analyse automatique de la structure prosodique d'énoncés de styles variés

Philippe Martin

philippe.martin@linguist.univ-paris-diderot.fr
LLF UMR 7110, Univ Paris Diderot Sorbonne Paris Cité

Abstract

A simple method for automatic prosodic structure analysis in French is proposed, operating from the identification of stressed syllables and the corresponding melodic curves glissando values. Applied to 4 different recording extracted from the C-PROM French corpus, this analysis gives some indications pertaining to the style of the speakers as well.

1. Structures prosodiques

Le but de cette étude est d'examiner la distribution des tons de frontière des groupes prosodiques à la fois du point de vue de leurs catégories, de leurs réalisations acoustiques et de leur conformité avec une grammaire prosodique. Si cette catégorisation automatique de ces tons s'avère pertinente, on disposera alors d'un processus automatisable pour déterminer la structure prosodique.

1.1. La structure prosodique autosegmentale-métrique

L'acceptation dominante du concept de structure prosodique relevant de la théorie Autosegmentale-Métrique (AM) organise hiérarchiquement en un ou plusieurs niveaux les groupes accentuels (*Accent Phrase*, AP) censés contenir une unité lexicale de classe ouverte (verbe, adverbe, nom ou adjectif) autour duquel gravitent un ou plusieurs mots grammaticaux, de classe fermée (conjonctions, pronoms, prépositions, etc.). Ces AP sont pourvus d'un accent mélodique (*Pitch Accent* dans la terminologie AM).

Dans une structure prosodique complexe, un premier regroupement des AP constitue un syntagme intonatif intermédiaire (ip), terminé par un ton de frontière. Le regroupement de ces ip constitue un syntagme intonatif (IP), également terminé par un ton de frontière. Enfin, le regroupement des IP constitue l'entièreté de la structure prosodique (SP), terminée par un troisième type de ton de frontière, dès lors conclusif.

Une structure prosodique donnée ne comprend pas nécessairement tous ces niveaux de regroupement. On peut par exemple avoir une SP "plate", regroupant une énumération de AP en un seul niveau. Une SP à deux niveaux est également possible, regroupant des AP en IP, et des IP en SP. La SP à trois niveaux regroupe alors des AP en ip, des ip en IP, et finalement des IP en SP. C'est le cas général mentionné plus haut.

1.2 Structure prosodique en français

En toute généralité, les groupes accentuels portent un accent mélodique lexical dont la position dans la séquence de syllabes est définie par la morphologie ou la syntaxe, ou encore par une règle rythmique. En français par contre, il n'y a pas d'accent lexical mais seulement un accent de groupe. On est donc conduit à admettre l'existence d'un troisième type de ton de frontière, placé sur la dernière syllabe prononcée des AP comme les tons de frontières de la SP, des IP et des ip.

De ce fait, il ne peut donc pas y avoir de AP en français, mais seulement des mots

prosodiques (*Prosodic Words*, PW). Ceux-ci n'étant pas des AP ne contiennent pas non plus nécessairement un seul mot lexical (Verbe, Adverbe, Adjectif ou Nom), mais peuvent en contenir plusieurs, ou encore se limiter à une seule syllabe.

Il y aurait donc en français 4 types de tons de frontières (et aucun accent lexical). Appelons les C0 (frontière de SP), C1 (frontière de IP), C2 (frontière de ip), et Cn (frontière des AP). Dans ce système, les mots prosodiques sont donc dotés en français d'un seul ton de frontière, qui peut être C0, C1, C2 ou Cn. Il n'y a pas de coexistence d'accent lexical et de ton de frontière comme cela peut être le cas en Italien par exemple. En fait, on rejoint par ce raisonnement la définition de la structure prosodique donnée il y a longtemps déjà par Martin (1975) et par Mertens (1987). Cette structure est clairement récursive en français.

1.3 Nature des tons de frontière

Ayant adopté le système de transcription ToBI, les descriptions de la SP autosegmentale-métrique utilisent des tons Haut et Bas (et leur variantes) pour décrire les différents tons de frontière. Cependant, l'existence manifeste de contrastes d'empan mélodique (contrastant les tons C0 et C2 par exemple) conduisent à utiliser plutôt des contours, dont les traits descriptifs impliquent par exemple la durée, la fréquence moyenne, et la variation mélodique. Outre des propositions de révision de la notation ToBI adaptée au français (Post & Delais, 2011), on trouve ainsi dans la littérature des définitions de tons de frontière par des contours -Haut, -Montant, -Ample pour C0, +Haut, +Montant, +Ample pour C1, +Haut, -Montant, -Ample pour C2.

Pas plus que la notation ToBI par tons Haut et Bas, ce dernier type de notation ne permet de rendre compte efficacement des réalités acoustiques manifestant ces différentes réalisations des tons de frontière, et en particulier du mécanisme dit du *contraste de pente*. Ce mécanisme prévoit essentiellement la réalisation pour C2 d'une variation mélodique de sens opposé à celle instanciée pour C1, dont C2 dépend (la relation de dépendance résulte du regroupement de ip terminés par C2 en un IP terminé par C1, la présence de C2 requérant celle de C1 à sa droite, i.e. apparaissant après C2).

1.4 Domaines des tons de frontière

Si l'analyse de corpus lus et fabriqués permet de valider expérimentalement les caractéristiques attendues des tons de frontière, il n'en va pas a priori de même avec la parole spontanée (i.e. non préparée). Toutefois, on peut s'attendre que, du point de vue de l'auditeur, les tons de frontière de même niveau présentent dans leurs réalisations suffisamment de traits communs pour qu'ils puissent être identifiés comme appartenant à une même classe, et ce au cours du déroulement du temps où surviennent les évènements prosodiques.

Contrairement à ce que peuvent suggérer les représentations graphiques des structures prosodiques, cette identification se fait séquentiellement dans l'axe temporel. Étant donné les limitations de la mémoire à court terme de l'auditeur, l'appartenance à une classe de tons de frontière donnée ne pourra se faire que relativement aux tons précédant et suivant le ton considéré. L'identification d'un ton est donc un processus local, par lequel l'auditeur doit décider si un ton (i.e. le contour mélodique qui le réalise) appartient à la même classe que le précédent ou non, ou éventuellement de plusieurs tons précédents.

Une exception à cette règle est donnée par le contour conclusif C0, dont les réalisations successives peuvent être temporellement éloignées, mais qui peuvent toujours être identifiées par les auditeurs quelles que soient ses variantes de

réalisation (l'auditeur peut toujours savoir si l'énoncé est terminé ou pas).

Du point de vue phonologique, cet aspect temporel revient à considérer l'existence de domaines locaux, définis comme des séquences de tons de frontière terminés par un ton de niveau supérieur. Dans chaque séquence de SP, l'auditeur devrait donc pouvoir identifier des contours terminaux C0 par des caractéristiques acoustiques similaires. De même, dans chaque séquence de IP, l'auditeur devrait pouvoir identifier des contours terminaux C1 par des caractéristiques acoustiques similaires, etc. Il en résulte que des contours de même classe C1, C2, Cn ne présenteront des traits semblables qu'à l'intérieur de chaque domaine, et ne seront pas nécessairement similaires du point de vue acoustique d'un domaine à l'autre dans l'énoncé.

Font également exception à ce processus les contours relatifs à l'accent secondaire (aussi appelé accent d'insistance), identifiés par des traits acoustiques similaires (montée mélodique), et qui, n'étant pas des tons de frontière, sont placés sur la première syllabe des mots lexicaux. Une ambiguïté apparait alors lorsque le mot lexical impliqué ne possède qu'une seule syllabe : s'agit-il alors d'un ton de frontière (éventuellement neutralisé) ou d'un accent d'insistance ?

Les catégories et les définitions fonctionnelles et perceptives des tons de frontière suivants :

C0 : ton de frontière conclusif, facile à identifier par simple écoute, au cours de laquelle l'auditeur n'attend pas de continuation de l'énoncé ;

C1, C2, Cn : catégories jugées perceptivement comme non conclusives. À l'écoute de segments extraits de l'énoncé et se terminant par un de ces contours, l'auditeur s'attend à une continuation de l'énoncé ;

Ci : correspondant à l'accent d'insistance, identifiable par sa position s'il n'est pas placé sur la syllabe finale des mots lexicaux ;

Creak : généralement utilisé par certains locuteurs comme ton conclusif, ou comme contour C1.

2. Analyse expérimentale

2.1. Corpus d'analyse

Si le mécanisme de neutralisation et la notion de domaines permet de mieux comprendre les variations de réalisations de tons de frontière observées, il pose aussi un problème difficile quant à la catégorisation des tons des données expérimentales. De plus, bien des auteurs ont remarqué que des tons apparemment classés de catégorie C1 étaient réalisés de manière très diverses.

Le corpus utilisé pour les tests est C-PROM (2010). C-PROM est un corpus aligné et annoté, développé pour l'étude des proéminences syllabiques en français. Il inclut 24 enregistrements échantillonnés en 7 genres (ou styles) de parole et produits par des locuteurs francophones (issus de Belgique, de France et de Suisse). Dans ce corpus, on n'a retenu que les locuteurs hexagonaux.

Les enregistrements du corpus C-PROM analysés appartiennent aux genres suivants :
lec-fr : lecture orale (149s.);
cnf-fr : conférence universitaire (224s.);
nar-fr : narration, récit de vie (197s.);
pol-fr : discours politique (217s.).

Le détail des formats de transcription peut être consulté en ligne.

2.2. Méthode d'analyse

À partir des transcriptions et des annotations de proéminences disponibles dans le corpus C-PROM, on a analysé plusieurs extraits de parole de locuteurs hexagonaux de styles différents (lecture orale, conférence universitaire, narration, discours politique). La durée totale des extraits est de 787s.

Dans chacun des extraits, les voyelles (et seulement les voyelles) des segments annotés comme proéminents dans les

annotations d'origine (proéminence forte P ou faible p, ce jugement étant éventuellement revu pour tenir compte de la contrainte des 7 syllabes qui limite le nombre de syllabes non proéminentes successives), ont été annotés et surlignés automatiquement.

Pour différencier C0, C1, C2 et Cn, on utilise comme critère la valeur correspondante de glissando en demi-tons par seconde (Mertens, 2004), du reste également affichée automatiquement par le logiciel d'analyse pour chaque segment. Si le glissando et supérieur au seuil de perception (n demi-tons par seconde au carré, avec une valeur du paramètre multiplicatif égale à 32), un contour mélodique montant réalisant le ton de frontière est catégorisé comme C1 ; S'il est descendant comme C2. Si le glissando du contour est inférieur au seuil, il est noté Cn.

Les segments ainsi annotés apparaissent sur l'écran d'analyse acoustique dans une couleur dénotant leur catégorie (Fig. 2).

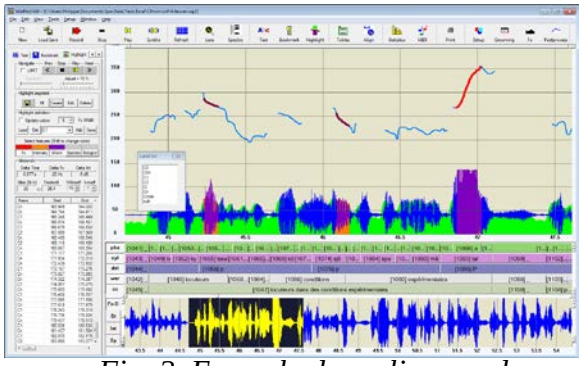


Fig. 2. Exemple de surlignage de segments en couleur différentes selon les catégories de contours prosodiques

Une fonction du logiciel d'analyse permet alors de mesurer automatiquement pour chaque segment correspondant à des tons de frontière annotés comme proéminents plusieurs paramètres acoustiques tels que la différence de fréquence fondamentale F0, la durée, la fréquence fondamentale moyenne, la valeur du glissando ainsi que le seuil correspondant. Ces valeurs sont automatiquement transférées dans un tableau Excel aux fins d'analyse. Ce logiciel dispose de plusieurs fonctions permettant l'analyse statistiques et des représentations efficaces des données, en particulier par l'utilisation des fonctions *powerview*. Par manque de place, ces résultats ne sont pas reportés ici.

2.3. Vérification de la grammaire prosodique

On a vu plus haut que les limitations de la mémoire à court terme de l'auditeur permettent de rendre compte du caractère local de la catégorisation des évènements prosodiques. Lors du déroulement temporel des évènements prosodiques, l'identification des contours prosodiques se fait domaine par domaine, chaque domaine étant défini par un contour de rang supérieur placé en fin de domaine : domaine des C0, contenant une séquence de C1, lesquels définissent des séquences de C1, les C1 terminant des séquences de C2, et les C2 des séquences de contours neutralisés Cn.

Cette hiérarchie de domaines permet de définit des séquences de classes de contours bien formées et mal formées. Ainsi, les séquences C1 C1 C0, C2 C1 C0, C1 Cn C0 sont bien formées, mais C1 C2 C0 ne l'est pas. De même C2 C2 C1, Cn Cn C1, Cn Cn C2 sont bien formées, mais pas C2 Cn C0.

La table 1 donne les occurrences de quelques-unes de ces séquences pour les 4 locuteurs. On constate qu'il n'y a qu'un seul cas de séquence mal formée ne correspondant pas aux principes donnés plus haut.

	lec-fr	**cnf-fr**	**nar-fr**	**pol-fr**
C1C0	2	2	1	3
*C2C0	0	0	0	1
C1C1	19	6	15	7
C2C1	7	2	1	9

Table 1. Distribution des séquences de contour selon le genre des locuteurs

La seule occurrence mal formée C2C0 apparait dans pol-fr, et à l'écoute s'interprète comme un accent d'insistance placé sur la syllabe finale du mot prosodique ce contour étant inattendu sur cette syllabe.

Le contraste de pente C2C1 s'observe relativement fréquemment dans cnf-fr, mais dans des séquences CnCn…CnC2C1, indiquant une structure prosodique [[[Cn Cn Cn] C2] C1] non congruente avec la structure syntaxique associée, alors qu'on s'attendrait à une SP [[Cn Cn Cn] C1] ou [[C2 C2 C2] C1]. La locutrice réalise donc un contraste de pente, mais seulement en précédant directement un contour C1, dénotant ainsi une moindre planification de la structure prosodique.

	lec-fr	**cnf-fr**	**nar-fr**	**pol-fr**
C0	7%	1%	1%	7%
C1	43%	25%	35%	38%
C2	9%	3%	1%	10%
Cn	30%	66%	53%	48%
Ci	2%	3%	4%	5%
creak	0.7%	0%	3%	0%
euh	0%	0.5%	4%	0%
Total	136	211	176	169

Table 2 Répartition des contours réalisés par genre

La table 2 présente les pourcentages d'emploi des différents contours, reflétant l'utilisant de structures prosodiques plus ou moins complexes. Ainsi les enregistrements lec-fr et pol-fr présentent un plus grand nombre d'occurrences de contours C2, donc de SP à 3 niveaux, caractéristiques de la parole lue (le locuteur pol-fr lit son discours). À l'inverse, la narratrice nar-fr utilise des structures plus simples, accompagnée d'un plus grand nombre d'hésitations.

3. Conclusion et perspectives

Le processus d'analyse présenté constitue une technique simple pour établir automatiquement la structure prosodique à partir de l'identification des syllabes effectivement accentuées. L'identification de ces syllabes utilise comme vérification la contrainte des 7 syllabes et la position finale des proéminences sur les mots lexicaux. Ce type d'analyse a déjà été tenté par des processus comme Analor (2013), mais seulement pour l'analyse de SP à un seul niveau.

L'identification des catégories de contour suppose, on l'a vu, la quasi linéarité de la variation mélodique des contours implicite dans le calcul du glissando. Elle dépend aussi de la valeur des paramètres retenus pour ce calcul (coefficient de variation d'intensité et coefficient de la différence de demi-tons). L'analyse de variations régionales ou idiosyncratiques présentant des contours convexes ou concaves impliquera l'élaboration de critères d'identification des contours plus élaborés, mettant en jeu les propriétés de contrastes locaux décrits plus haut.

Références

Analor (2013). Logiciel d'étiquetage et séquencage basée sur l'analyse Prosodique du discours, http://www.lattice.cnrs.fr/Analor,70

C-PROM (2010). Corpus libre de parole multigenre, https://sites.google.com/site/corpusprom/

Martin, Ph. (1975). Analyse phonologique de la phrase français. *Linguistics* 146, pp. 35-68.

Martin, Ph. (1987). Prosodic and Rhythmic Structures in French *Linguistics* 25:5, pp. 925-949.

Mertens, P. (1987). *L'intonation du français. De la description linguistique à la reconnaissance automatique.* Unpublished Ph.D. (Univ. Leuven, Belgium).

Mertens, P. (2004). Le prosogramme : une transcription semi-automatique de la prosodie *Cahiers de l'Institut de Linguistique de Louvain* 30, 1-3, pp. 7-25.

Post, B. and Delais-Roussarie, E. (2011). French ToBI, Workshop on Romance ToBI, *Satellite workshop PaPI 2011*, Universitat Rovira i Virgili (Tarragona).

Towards automatic detection of prosodic boundaries in spoken French

***Piet Mertens*° *& Anne Catherine Simon*°°**

Piet.Mertens@arts.kuleuven.be, Anne-Catherine.Simon@uclouvain.be
°Linguistics Department, University of Leuven (KU Leuven), Belgium
°°Institut Langage & Communication / Valibel, Université catholique de Louvain, Belgium

Abstract

This paper describes a tool for automatic detection of prosodic boundaries (PBs) in French, and evaluates it on a 12 min. speech corpus, using a reference annotation prepared by a trained phonetician.

The phonetic alignment of the corpus is used to identify the nucleus (as the local peak of intensity) within the rhyme of each syllable. After pitch stylization, the following prosodic properties are computed for each syllable: pause presence, syllable lengthening, intra-syllabic pitch movement, and pitch prominence. These features are combined in detection rules. No training corpus is required.

Perceived PBs of 4 strengths (major, intermediate, minor, no boundary) were annotated by an expert, for a test corpus of 8 samples of continuous speech.

The match between reference PBs and detected PBs was evaluated as a classification task. In addition, the contribution of each prosodic feature to boundary detection was estimated.

1. Introduction

One of the functions of prosody is to divide speech into chunks of one or more words. The term *prosodic boundary* (PB) is used in two ways, to refer either to the limits of these chunks, or to speech properties that mark these limits. According to Ladd (2008: 288), PBs are remarkably difficult to define and to identify consistently and as a result, there is often disagreement about whether a PB is or is not present.

Prosodic units and PBs are relevant for spoken language research and discourse analysis, where segmentation is central for the analysis of turn construction in conversation (Selting 2000), the prosody-syntax interface (Mertens 2006), discourse relations (Wichmann 2000), the understanding of discourse production and processing (Clifton *et al*. 2006), as well as for speech recognition and text-to-speech synthesis.

The *acoustic and perceptual cues* (Wagner & Watson 2010: 907-910) of PBs are related to the presence of a pause, to duration (pre-boundary lengthening, domain-initial strengthening), pitch (pitch movement, pitch discontinuity at the PB, including declination line reset, and the relative height of pitch accents), to intensity, and phonation type (creak).

In prosody research PBs are characterized in various ways (Wagner & Watson 2010: 911). The *qualitative* view assumes a small set of boundary *types*, associated with the corresponding *hierarchical prosodic units* (such as phonological word, intonational phrase), or with functionally defined *boundary strengths* (non-terminal vs. terminal PB, or continuation vs. final PB). In a second, *quantitative* view, the strength of a PB is expressed *relative* to the strength of boundaries occurring earlier in the utterance, without reference to prosodic units or functions. In a third view, boundary strength is approached as a *perceptual* notion in its own right, without reference to prosodic structure, and which can be measured directly in perception experiments with untrained listeners (Pijper & Sanderman 1995). In this paper a PB is defined as a perceived separation between two chunks of speech, marked by prosodic means such as pause, pitch movement, pitch discontinuities and lengthening.

In free stress languages the stressed

syllable and the syllable preceding the PB are often dissociated. As a result, acoustic correlates of PBs may be associated with the pre-boundary syllable(s), or with events (pause, pitch discontinuity…) at the PB itself. Instead, in some fixed stress languages, stress usually occurs at a PB, in which case both stress and boundary features are associated with the same syllable. This is the case in French, where word stress is on the final syllable of an intonation unit (Mertens 1993; Di Cristo 1999).

The variety of PB definitions raises the question whether PBs may be *identified consistently* by human annotators. In manual labelling by trained listeners, based on perception, judges may rely on their ideas about prosodic organization and use the segmental information to identify syntactic structure (Campbell, 1993: 10; Pijper & Sanderman 1995: 2038). However, perception experiments (Pijper & Sanderman 1995) suggest that untrained listeners can give reliable judgments of PB strength, even when the lexical contents of the utterances is made unrecognizable ("delexicalised").

How many *levels of boundary strength* should be distinguished? In perception experiments where listeners are asked to rate PBs on a 10-point scale, the number of resulting PB categories is usually lower (Pijper & Sanderman 1995). Grover *et al.* (1997) suggests that boundary strength is transcribed more consistently using a 4-value scale than using a scale with a many values. But often the number depends on theoretical assumptions about PBs. Phonological models of French intonation (e.g. Martin 1978, Rossi 1999, Mertens 2006) often imply three or more levels of boundaries (cf. minor and major continuation, and terminal contour of Delattre 1966). Annotated corpora commonly use two or three levels: the C-ORAL-ROM corpus annotation (Moneglia *et al*. 2005) distinguishes terminal and non-terminal PBs.

Automatic detection of PBs relies on their acoustic cues (cf. supra), although the approaches and algorithms for the actual detection differ considerably (Ostendorf 2000). The experiments of Pijper & Sanderman (1995), using judgments by untrained listeners, suggest that, for Dutch read speech, the most important cues of PBs are the presence of a pause (> 200ms), the melodic discontinuity (drop of pitch during the silent PB), the declination line reset and pre-boundary lengthening. The relation between PBs and temporal organisation (duration of phonetic segments in syllable onset, nucleus and coda) is studied in detail in Campbell (1993, 2000). There have been efforts to combine into a single algorithm the detection of pauses, lengthening, F0 variation, prominence, in order to segment spoken French into major prosodic units (Lacheret-Dujour & Victorri 2002), also using lexical information.

2. The system for PB detection in French

Our strategy strongly relies on the prosodic structure of French, in which the last syllable of the intonation unit may carry particular pitch movements, may be prominent for pitch or duration (lengthening), and followed by a pause.

2.1. Prosodic features measured

The following properties are measured: pause presence, syllable lengthening, intra-syllabic pitch movement, and pitch prominence.

A syllable is *prominent for some prosodic attribute* (duration, pitch) when it stands out from its context due to a local difference for that attribute (Mertens 1991). Prominence may be quantified as the value at the target syllable, divided by the mean value in the context. Depending on the number of syllables in the left and right context, the context window may be symmetrical or asymmetrical, fixed or dynamic; in the latter case, window length depends upon the properties of the syllables in the context.

The *phonetic alignment* provides the syllables, vowels and rhymes. The *syllabic nucleus* is determined as the voiced part of the rhyme (= vowel + coda), located around the intensity peak of the vowel, for which the intensity drop stays below some threshold and provided it does not include pitch discontinuities.

A *pause* is detected when the interval between successive nuclei exceeds 200ms.

Hesitations ("euh" vowels in French) affect the estimation of lengthening and pitch prominence, more generally when they appear in the context used for measuring prominence. A hesitation is detected when a syllable is labelled [œ], [ø] or [ə], has a duration of at least 350 ms and a pitch which is level to slightly falling (> -3 ST).

Syllable lengthening is measured as syllable duration prominence (*SDP* = syllable duration / mean syllable duration in context window), for an asymmetrical and dynamic context window of at most 2 syllables to the left and 1 to the right.

The following problems were encountered. (1) Hesitations are most often considerably longer than other syllables, affecting syllable length measurement. (2) Pauses act as perceptual boundaries for the context window. Therefore, the context window is truncated at a pause, and its width is adapted dynamically (max. 500ms for each side). To avoid artefacts due to small context size, *SDP* is set to 1 (hence, no lengthening) when the context contains only 2 syllables. (3) At high speech rate, intrinsic duration of speech sounds largely affects sound duration. To avoid this, *SDP* is set to 1 for nuclei of 40 ms or shorter.

Pitch prominence is measured as prominence of the mean pitch value (the mean F0 within the syllabic nucleus), for a context size of 2 syllables to the left and 1 syllables to the right, using a dynamic context width.

The values for intra-syllabic *pitch rise* and *pitch fall* indicate the cumulated positive, resp. negative, pitch intervals within the syllabic nucleus.

2.2. Rules for PB assignment

The rules for PB assignment given below are based on empirical observation of corpus data annotated by a phonetician. They are similar to those of other studies (e.g. Lacheret & Victorri 2002).

1. Do not assign a PB to hesitation syllables (which are detected automatically).
2. Assign a *major PB* (level 3) in three cases: (a) when duration prominence (SDP) exceeds 3 (i.e. 3 times as long as the context mean, corrected for high speech rate), (b) when the nucleus contains a pitch rise or a pitch fall of 10 ST or more, or (c) when it is followed by a pause of at least 200 ms.
3. Assign an *intermediate PB* (level 2) in three cases: (a) when SDP exceeds 2 (corrected for high speech rate), (b) when the pitch rise in the nucleus is at least 4 ST, or (c) when pitch prominence is 5 ST or more.
4. Assign a *minor boundary* (level 1) when SDP exceeds 1.5, provided nucleus duration is at least 40 ms.

3. Evaluation

3.1 Speech material

The test corpus contains 8 speech samples[1] of approx. 100s, by 9 speakers, male and female, with a total duration of 737s (3029 syllables). Four samples consists of unprepared speech (radio-interviews, conversations), the other four of read-aloud speech (radio-news, conference presentations). A

[1] All samples, except the directions request, are taken from the Valibel Speech Database (Dister *et al.* 2009) and illustrate a (standard) Belgian variety of French. Samples under the category "academic" represent academic discourse at official occasions. Radio-news and radio-interview are broadcasts from the national radio programs. Interview comes from sociolinguistic investigation about standard usages of French. Directions request comes from a M. Avanzi corpus collected in France, and available in the C-Prom project (Avanzi *et al.* 2010). Finally, everyday conversation involves two close friends self-recorded at home.

validated phonetic alignment was prepared by the authors. Subcorpus A contains all 3029 syllables; in subcorpus B hesitations (detected or marked in the corpus annotation) and syllables without a detected syllabic nucleus (either because it was unvoiced, too short or contained pitch discontinuities), are discarded, resulting in a set of 2625 syllables.

3.2. Reference annotation

Manual labelling of the speech material was carried out by a trained phonetician, who is a native speaker of French. The annotator listened to sound fragments of 4 to 8s, played 3 times. For each word-final syllable (i.e. a potential stress), the annotator assigned one out of four boundary levels, either 0 (no PB), 1 (minor PB), 2 (intermediate PB) or 3 (major PB). Syllables with emphatic initial stress (ES) and hesitations were also identified by the annotator, but treated as "no PB" in the evaluation described here.

3.3. Results

The automatic PB detection was evaluated as a classification task, mapping *observed* categories to *actual* ones. In this case, the observed category is the automatically detected PB and the *actual* category is provided by the reference labelling of the phonetician.

For *minor PBs* (level 1) very poor results are obtained. In French, pitch contours are anchored at the final syllable of an intonation group, which coincides with the last syllable of a syntactic constituent, and PBs are very likely at the end of a syntactic constituent. As a result PB perception may be biased by segmental information about syntactic structure. This holds for PBs of all levels, of course, but for PBs of level 2 and 3 acoustic cues are usually present. In the evaluation, minor PBs were removed (i.e. replaced by "no PB") from the reference annotation and the detected PBs.

The confusion matrix for subcorpus B is shown in table 1. The detection of major PBs (level 3) is rather good, whereas that of intermediate PBs (level 2) is poor.

	observed			
actual	0	2	3	
0	2041	112	80	2233
2	69	81	24	174
3	11	16	191	218
	2121	209	295	2625

Table 1. Confusion matrix for subcorpus B. (observed = detected PB; actual = reference PB)

The system detects noticeably more level 2 PBs than the human expert (209 vs. 174). This is partly explained by the fact that the human annotator distinguishes between final stress (followed by a PB) and emphatic initial stress ("ES", treated as "no PB" in the evaluation data), whereas the system does not detect ES as such, and as a result many syllables carrying ES will be detected as a level 2 PB (but not as a level 3 PB, since ES is not followed by a pause).

80 syllables without a PB were detected as major PBs. This is explained by the fact that the last syllable of an utterance is often devoiced, in which case its nucleus is not detected, and no PB will be detected either. Also, the skipped syllable may be interpreted as a pause, resulting in a major PB being detected at the preceding syllable.

	PB	N	%	Prec.	Rec.	Accur.	F
A	0	2586	85.4	94.8	91.7	88.6	93.2
	2	182	6.0	37.7	44.5	92.2	40.8
	3	259	8.6	61.4	73.7	93.8	67.0
A'	0	2586	85.4	94.8	91.4	88.4	93.1
	2	182	6.0	37.6	45.1	92.2	41.0
	3	259	8.6	59.9	73.7	93.5	66.1
B	0	2233	85.1	95.7	93.8	91.1	94.7
	2	174	6.6	38.6	44.8	91.6	41.5
	3	218	8.3	79.0	84.4	96.8	81.6

Table 2. Classification results for the automatic detection of PBs using 3 boundary levels (0, 2 and 3) for subcorpora A, A' (see text) and B. (Data set does not include level 1 PBs.) N = count.

Table 2 shows precision, recall, accuracy and F-measure for each PB class (0, 2, 3), as well as the number (N) and percentage (%)

of elements in each class, for subcorpora A and B. The elimination of hesitations and syllables without a detected nucleus – in subcorpus B – slightly improves recall and precision, in particular for major PBs: 84.4% of the actual major PBs are indeed detected as major PBs (recall) and 79% of the detected major PBs are actual major PBs (precision). Syllables without a PB are detected with a precision of 95.7% and a recall of 93.8%.

Finally, results for "A'" correspond to subcorpus A, when decision rule 1 is disabled, i.e. when hesitations are treated in the same way as other syllables. The small difference between A and A' shows the impact of rule 1 is negligible.

Table 3 shows the contribution of individual prosodic cues to the detection of *actual* PBs of level 2 and 3, as the percentage of boundaries for which a given cue was present. In subcorpus B, for *major* PBs, pause is the most effective cue (85.8%), followed by lengthening (11%). For *intermediate* PBs the most important cues are lengthening (24.1%) and pitch prominence (24.1%), whereas the contribution of pause drops to 9.8%.

PB	N	P	R	F	r	T	L2	L3
2	174	9.8	0.6	1.7	6.9	24.1	24.1	2.3
3	218	85.8	5.5	1.4	4.6	6.0	1.4	11.0

Table 3. Contribution (percentage present) of prosodic cues to PB detection for PB levels 2 and 3, in subcorpus B. N=number of syllables, P=pause, R=large rise (≥10ST), F= large fall (≤10ST), r=small rise (≥4ST) , T=pitch prominence (≥5ST), L2=lengthening ≥2, L3=lengthening ≥3 .

Using pause as the *only* cue results in the detection of 280 *major* PBs, against 295 when *all* cues are used. Note that the number of observed (detected) PBs exceeds the number (218) of actual major PBs. Table 4 shows the classification results for *major* PBs only, when level 2 PBs are treated as "no PB". It shows the use of cues other than pause improves recall form 85.8% to 87.6%. The scores of tables 2 and 4 should not be compared, since they represent distinct classification tasks: 3 classes (0, 2, 3) for table 2, against 2 classes (0, 3) for table 4.

cues used	N actual	N obs.	Prec.	Rec.
pause only	218	280	66.8	85.8
all cues	218	295	64.7	87.6

Table 4. Classification results for subcorpus B for the automatic detection of major PBs using either pause only or all prosodic criteria.

4. Discussion

Common causes of errors may be identified. The first, syllable devoicing, occurs when one or more syllables at the end of an utterance are pronounced with gradual or complete devoicing, possibly with creak, for instance when sub-glottal pressure decreases or when the pitch drops to the bottom of the pitch range. Such unvoiced syllables will not be recognized as syllables by the algorithm, either due to low intensity, lack of voicing, or octave jumps typical of creak.

In French, the syllable with final stress may sometimes be followed by a schwa, detected as a separate syllable or even as a hesitation. This will affect the location of the detected PB.

In a third type of error a PB is detected at an initial emphatic stress. Currently the system is unable to distinguish the two types of stress found in French.

The fourth type of error concerns hesitations. The perception of hesitation is a complex phenomenon, which implies prosodic cues (lengthened syllables, flat or slightly falling pitch contour) but also syntactic phenomena (intra-phrase silent pauses, repetitions and false starts, cf. Duez 2001). Only 25% of the hesitations in the reference annotation are detected by the system.

5. Conclusion

Speech corpora are largely available today, but still require more reliable tools for annotation tasks, especially for prosodic annotation. In this contribution, we propose

an automatic tool for PB detection in spoken French. Detection is purely acoustic: it is based on a detection of syllable prominence (pitch movement or pitch peak, lengthening) in a local context, and pause. One out of three boundary levels is assigned, depending on the characteristics of the syllable. Good results are obtained for detection of major PBs and acceptable results for the detection of intermediate PBs.

Analysis of the results shows that the most important cue for *major* PBs is pause (85.8%), followed by lengthening (11%), whereas for *intermediate* PBs, the most effective cues are lengthening (24.1%) and pitch prominence (24.1%).

The analysis of frequent errors suggests improvements of the algorithm. First, access to lexical information (syllable position within the word, detection of repetition or hesitation particles) would help in interpreting prosodic variation (like syllable lengthening) that may fulfil very diverse functions, according to its location. Second, a better description of the acoustic properties of emphatic initial stress (ES) in French might provide us with tools for distinguishing final vs. initial stress.

References

Avanzi, M., Simon, A.C., Goldman, J.-P. & A. Auchlin. (2010). C-PROM. Un corpus de français parlé annoté pour l'étude des proéminences. *Actes des 23èmes JEP* (Mons, 25-28 mai 2010).

Campbell, W.N. (1993). Automatic detection of prosodic boundaries in speech. *Speech Communication* 13, pp.343-354.

Campbell, W.N. (2000). Timing in Speech: A Multi-Level Process. Horne, Merle (ed.) *Prosody: Theory and Experiment.* Dordrecht, Kluwer.

Clifton, C., Carlson, K. & Frazier, L. (2006). Tracking the what and why of speakers' choices: prosodic boundaries and the length of constituents. *Psychonomic Bulletin & Review* 13:5,pp. 854-61.

Delattre, P. (1966). Les dix intonations de base du français. *French Review* 40/1, pp.1-14.

de Pijper, J. R. & Sanderman, A. (1995). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *JASA* 96, pp. 2037-2047.

Di Cristo, A. (1999). Vers une modélisation de l'accentuation en français (première partie). *J. of French Language Studies* 9:2, pp. 143-163.

Dister, A., Francard, M., Hambye, Ph. & Simon, A.C. (2009). Du corpus à la banque de données. Du son, des textes et des métadonnées. L'évolution de banque de données textuelles orales VALIBEL (1989-2009). *Cahiers de Linguistique* 33:2, pp. 113-129.

Duez, D. (2001). Signification des hésitations dans la production et la perception de la parole spontanée. *Revue PArole* 17-18-19, pp. 113-137.

Grover, C., Heuft, B., Coile, B. van (1997). The reliability of labeling word prominence and prosodic boundary strength. *Proc. ESCA Workshop on Intonation*, Athens, pp. 165-168.

Lacheret, A. & Victorri, B. (2002). La période intonative comme unité d'analyse pour l'étude du français parlé: modélisation prosodique et enjeux linguistiques, *Verbum* XXIV/1-2, pp. 55-72.

Ladd, D.R. (2008). *Intonational Phonology*. Cambridge University Press, Cambridge.

Martin, Ph. (1978). Questions de phonosyntaxe et de phonosémantique en français. *Lingvisticae Investigationes* 2, pp. 93-126.

Mertens, P. (1991). Local prominence of acoustic and psychoacoustic functions and perceived stress in French. *Proceedings of the 12th International Congress of Phonetic Sciences* 3, pp. 218-221.

Mertens, P. (1993). Intonational grouping, boundaries, and syntactic structure in French. House, D. & P. Touati (eds.), *Proc. ESCA Workshop on Prosody* (Sept. 27-29, 1993, Lund), pp. 156-159.

Mertens, P. (2006). A Predictive Approach to the Analysis of Intonation in Discourse in French. In Kawaguchi, Y.; Fonagy, I.; Moriguchi, T. (eds.), *Prosody and Syntax*. John Benjamins, Amsterdam, pp. 64-101.

Moneglia M., Fabbri M., Quazza S., Panizza A., Danieli M., Garrido J., Swerts, M. (2005). Evaluation of consensus on the annotation of terminal and non-terminal prosodic breaks in the C-ORAL-ROM corpus. Moneglia, M. (ed.), *C-ORAL-ROM: integrated reference corpora for spoken Romance language*, Benjamins, pp. 257-276.

Ostendorf, M. (2000). Prosodic Boundary Detection. Horne, Merle (ed.), *Prosody: Theory and Experiment.* Kluwer, Dordrecht.

Rossi, M. (1999). *L'intonation, le système du français: description et modélisation*. Ophrys, Paris-Gap.

Selting, M. (2000). The construction of units in conversational talk. *Lang. in Society* 29, pp. 477-517.

Wagner, M. & Watson, D.G. (2010). Experimental

and theoretical advances in prosody: A review. *Language and Cognitive Processes* 25, pp. 905-945.

Wichmann, A. (2000). *Intonation in Text and Discourse. Beginnings, middles and ends*. Harlow, Longman.

The production of Dutch word stress by Francophone learners

Marie-Catherine Michaux[1], Johanneke Caspers[2]

marie-catherine.michaux@uclouvain.be, j.caspers@hum.leidenuniv.nl
[1]Université catholique de Louvain – F.R.S.-FNRS
[2]Leiden University Centre for Linguistics

Abstract

This study aims at exploring the production of Dutch word stress by Francophone learners of (Belgian) Dutch. Following other studies, it was hypothesized that participants would show a preference for stressing the final syllable. This hypothesis was confirmed, but the large variability in the data and the lack of agreement between labellers suggest that there is more to it.

1. Introduction

Dutch is a variable-stress language where stress is a lexical property of words (Rietveld & van Heuven 2009) that can be used contrastively (e.g., *voorkomen,* 'to happen', vs. *voorkomen,* 'to prevent'). French does not have contrastive stress: the standard final prominence in isolated words disappears when they are located in non-final position in a larger word group, leaving a word-group final accent (Lacheret-Dujour & Beaugendre 1999; Di Cristo 2000; Rasier 2006). Rather than being contrastive, this 'primary' accent has a demarcative function.

In Dutch, word stress is used as an important cue for word recognition (Cutler 2012; Van Leyden & van Heuven 1996) and Dutch speakers have been shown to be sensitive to mis-stressing (Cutler 2012).

Because French speakers do not have a linguistically encoded prominence at the word level they have sometimes been claimed to be 'stress-deaf' (e.g., Peperkamp & Dupoux 2002; Altmann 2006). However, recent research seems to show that with training Francophones might be able to perceive stress contrasts (Schwab & Llisterri 2012).[1]

In French-speaking Belgium, Dutch is taught as a foreign language in most primary and secondary schools. According to the surveyed students and teachers, pronunciation and prosody, however, are often neglected in Dutch as a Foreign Language (DFL) courses, so that most learners may not be familiar with Dutch word stress.

The production of Dutch word stress has been addressed in small-scaled studies with Francophone learners of Dutch as a second language (DSL) by Caspers and van Santen (2006) and as a foreign language (DFL) by Heiderscheidt and Hiligsmann (2000) and Michaux et al. (2012). Based on the results of these studies it seems clear that the DFL population has to be analysed separately from the DSL one, as the latter group, probably as a result of receiving another type of input (viz. native spoken Dutch), has been found to be more proficient in producing correctly located stress. As for the DFL group it was concluded that learners tend to stick to their L1 pattern, but can also evolve to a penultimate stress (yet not always being the required stress position in Dutch) as time goes by.

This paper investigates DFL word stress production. Following Caspers & van Santen (2006), Michaux et al. (2012) and Schwab & Llisterri (2012), we hypothesized

[1] To our knowledge, no study has dealt with the perception of Dutch word stress by Francophone DFL learners, which will be the following step in the current research project.

that Francophone speakers will transfer their L1 final prominence pattern to Dutch words. As familiarity with a phrase-final accent in the L1 might lead to a bias towards final prominence in that position, we also investigated this hypothesis. As a result, non-phrase final words might bear final stress less often in DFL production than phrase-final ones.

2. Method

2.1. Participants

20 DFL learners (age range 19-23, mean age 21.1, 14 females) and 10 native speakers of (Belgian) Dutch (age range 20-51, mean age 28.6, 5 females) took part in the experiment. French was the only mother tongue of the selected DFL speakers.

2.2. Materials

30 existing Dutch three-syllable words were used. They were classified according to the stress rules for simplex words by Trommelen & Zonneveld (1989). The words were split into three canonical stress positions: initial (*pagina,* 'page'), medial (*collega,* 'colleague') and final (*anoniem,* 'anonymous'). Each word was presented thrice in a carrier sentence (*X heb ik gezegd* 'X I said', *Ik heb X gezegd* 'I X said' and *Ik heb gezegd X* 'I said X'), leading to a 90-sentence reading task.

2.3. Procedure

Speakers were recorded individually in a quiet room. Before the recording they filled in a form containing questions about their learner profile (length of Dutch learning, age at start of learning), age and other known/spoken languages.

The trial phase started after an instruction and training session similar to the trial. A Tascam-07 MKII recorder and a Sennheiser PC131 head-set microphone were used.

2.4. Analysis

The data were perceptually labelled independently by a highly-proficient Dutch speaking native French speaker (Labeller 1, Lab 1) and two native Dutch speakers (Lab 2, Lab 3), all of whom were phonetically trained. After listening as often as required to the stimuli, the labellers indicated which syllable they perceived as stressed (1-2-3). Cases of doubt could be expressed as "1?3?", etc.

3. Results

3.1. Interrater agreement

Interrater reliability on the labelled DFL data yielded $\kappa = 0.772$ (Lab 1 and Lab 2), $\kappa = 0.674$ (Lab 1 and Lab 3) and $\kappa = 0.684$ (Lab 2 and Lab 3), all agreements thus being 'substantial' according to Landis & Koch (1977), but not perfect. The agreement between Lab 1 and 2 is slightly higher than the other combinations. The highest κ-scores per participant are found for speaker DFL23 ($\kappa = 0.987$, 0.983 and 0.987 resp.), whereas the lowest are found for DFL22 ($\kappa = 0.128$, 0.132 and 0.405). Low kappas indicate that the labellers generally did not hear and mark the same syllable as prominent or that they hesitated between several syllables.

For the native speaker group interrater agreement is almost perfect: $\kappa > 0.98$ for all pairs of raters.

		No consensus		Syll 1		Syll 2		Syll 3	
Canonical SP	1	27.0	(162)	15.8	(95)	16.5	(99)	40.7	(244)
	2	26.3	(158)	7.2	(43)	30.2	(181)	36.3	(218)
	3	17.4	(104)	9.5	(57)	8.0	(48)	65.1	(390)
		23.6	(424)	10.8	(195)	18.2	(328)	47.4	(852)

Table 1: Percentages (and counts) consensus between labellers for 1st, 2nd and 3rd canonical stress position broken down by perceived stress position.

3.2. Consensus

Based on the labels per labeller, a consensus variable was computed, consensus being reached when per word all labellers marked the same syllable as prominent.

Table 1 shows the consensus values per canonical stress position (SP) for the DFL speakers. The shaded cells contain the cases where canonical and perceived stress concur, leading to correct results. The overall percentage of correct stress amounts to 37.0% (vs. 97.9% for the native group), see 3.3. for more details. According to the hypothesis, canonical SP3 should have been least problematic, which has been born out: SP3 yields the best results (65.1% correct), followed by SP2 (30.2%), which is at chance level, and SP1 (15.8%). On the whole there is a preference for syllable 3 (47.4%) regardless of the canonical SP, but there is also substantial variation in the remainder of the data. While there is a clearer preference for syllable 3 in the other cases, SP2 shows a different pattern: both syllable 2 and 3 yield approx. 30% of consensus (chance).

In roughly 25% of the cases with SP1 and SP2, no agreement (i.e. 'no consensus' in the table) was reached between the labellers, while this percentage reaches 17.4% for SP3. Strikingly, 'no consensus' very often has the second highest frequency after consensus on the 3rd syllable. This either means that the labellers could not determine which syllable was stressed or that they perceived different or several syllables as stressed.

3.3. Percentage of correct stress

The percentage of correct stress in the DFL material being 37.0% (see 3.2.), it seems safe to claim that the DFL speakers in this study had not mastered Dutch word stress yet.

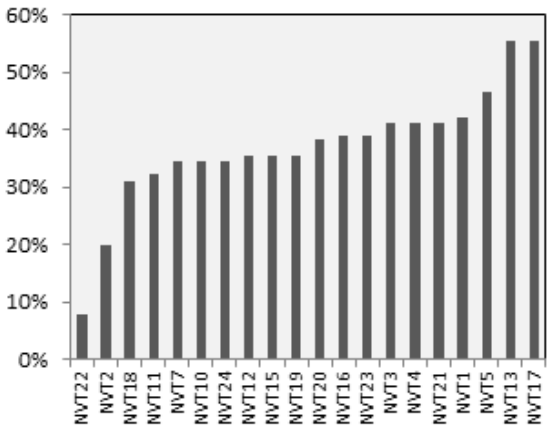


Figure 1: Percentages of correct stress per speaker, ranked from lowest to highest.

Figure 1 shows the percentage of correct word stress per speaker. The highest scores are reached by DFL13 and DFL17, but do not exceed 60.0%. The lowest score is found for DFL22 who unsurprisingly is also the participant with the lowest κ-values. However, the participant with the highest κ-scores (DFL23) does not achieve the highest correctness score (38.9%), meaning that a clear stress realisation does not necessarily imply a correct stress location. The majority of the participants reach between 30% and 40% correctness, which again shows how little grasp the speakers had of Dutch word stress.

A repeated measures ANOVA (with Greenhouse Geisser correction) of the percentage correct, aggregated over stimulus words, with canonical SP and word position in the sentence as within-subjects factors and L1 as between-subjects factor, shows an effect of L1, $F_{(1,28)}= 324.1$ ($p<.001$) and canonical SP, $F_{(1.6,45.3)}= 8.5$ ($p<.001$), and an interaction between canonical SP and L1 ($F_{(1.6,45.3)}= 8.3$ ($p<.001$). This means that the DFL production varies a lot more for different canonical SPs than the native production does. Pairwise comparisons reveal that the effect of SP is caused by the difference between SP3 and the other SPs. Contrary to our hypothesis, there is no effect of the position in the sentence ($F_{(2.0,55.0)}=$

2.2, ins.), meaning that there is a bias towards syllable 3 at the word level (see 3.2.), but not at the phrase level. There is no interaction between L1 and position in sentence ($F_{(2.0,55.0)}$= 1.6, ins.), and between position in sentence, L1 and SP ($F_{(3.4,95.4)}$< 1, ins.).

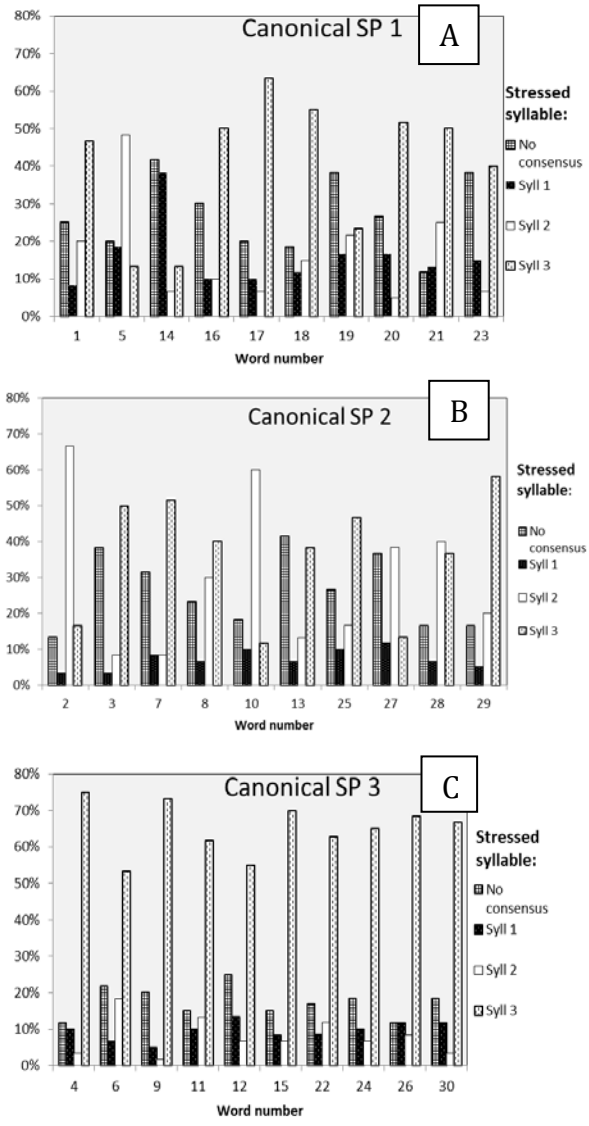


Figure 2: Stressed syllable per word and canonical SP1 (A), SP2 (B) and SP3 (C).

3.4. Variability in the data

Figure 2 shows the stressed syllable (based on consensus) per canonical SP and stimulus word. SP3 (panel C) shows less variability ($\chi^2_{(27)}$= 31.24, ins.) than categories SP1 (panel A) and SP2 (panel B) where the stressed syllable seems to vary more depending on the stimulus word pronounced (resp. $\chi^2_{(27)}$= 144.8, p<.001 and $\chi^2_{(27)}$= 147.5, p<.001).

Given the variation in the data, the words were further analysed according to their form and stress pattern similarity to French. The words were split into four categories: (1) same form, same prominence location, e.g., *chocola*, Fr. 'chocolat' ('chocolate'), (2) same form, different prominence location, e.g., *marathon*, (3) different form, same prominence location, e.g., *abrikoos*, Fr. 'abricot' ('apricot'), (4) different form, different prominence location, e.g., augustus, Fr. 'août' ('August'). Figure 3 shows the distribution of stressed syllables per category.

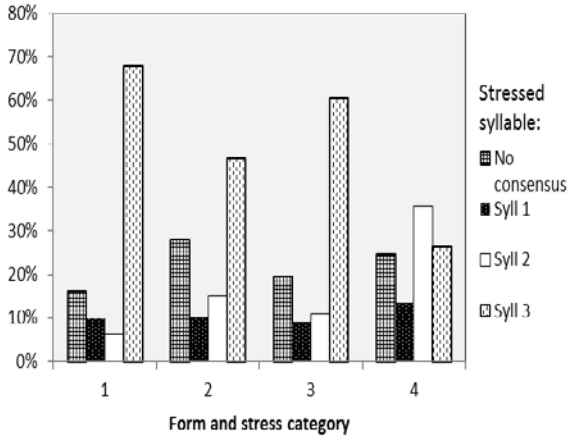


Figure 3: Percentage of stressed syllable per form and stress category (see text).

Categories (2) and (4) yield proportionally more 'no consensus' (resp. 28.1%, 24.6%) than other categories. Furthermore, category (4) yields stress more often on syllable 2 (35.6%) than other categories ((1) 6.7%, (2) 15.1%, (3) 10.0%). In contrast, stress on syllable 3 occurs much less often in category (4) (26.3%) than in other cases ((1) 67.8%, (2) 46.7%, (3) 60.5%). Category (4) comprises words such as *collega* (Fr. 'collègue', En. 'colleague') and *augustus*. One could argue that a correct stress location on the second syllable is related to

the high frequency of those words. If anything, this result also suggests that words formally less similar to French should be viewed separately, as they seem to lead to another stressing behaviour and might be less prone to the transfer of the French final pattern.

4. Conclusion and discussion

The low scores of the participants clearly show poor knowledge of Dutch word stress. On the whole, the speakers relied on their final L1 pattern to stress the stimuli regardless of the L2 canonical stress position. This globally supports our hypothesis.

However, all cases taken together, the highest percentage of final stress amounted to 47.7%, which means that in the remainder of the cases there is either variability in the realized stress position or lack of agreement between the labellers. The second most frequent labelling after syllable 3 was not any other stress position, but the category 'no consensus'; after the labelling process the labellers all stated how unstable they had found the data, some of the stimuli bearing several prominences, sometimes felt to be rendered with different cues. In the aftermath of this production study all production data will be acoustically analysed to find the reason for this perceptual uncertainty. Hypothetically, if several syllables were indeed stressed, one could argue that some kind of 'arc accentuel' has been produced (Di Cristo 1998) or that the participants showed an unsteady behaviour because of the lack of a mentally stored stress. However, as some words (category (4)) showed a preference towards stress on the second syllable, one could also argue that in some cases the participants realised that there is a prosodic difference between Dutch and French. This would be in agreement with the findings of Michaux et al. (2012).

Finally, contrary to our expectations, a final location of the word within the sentence did not yield final stress more often, which forces us to reject our second hypothesis.

Acknowledgments

The first author is supported by a grant from the Fonds National de la Recherche Scientifique (F.R.S-FNRS). We would like to thank Prof. V.J. van Heuven, Prof. Ph. Hiligsmann and Prof. L. Rasier for comments on an earlier version of this article.

References

Altmann, H. (2006). *The Perception and Production of Second Language Stress: A Crosslinguistic Experimental Study*. Diss., University of Delaware.

Caspers, J. & A. Van Santen (2006). Nederlands uit Franse en Chinese mond. Invloed van T1 op de plaatsing van klemtoon in Nederlands als tweede taal?[Dutch by French and Chinese learners. L1 influence on stress position in Dutch as a second language] *Nederlandse Taalkunde, 11*(4), pp. 289-318.

Cutler, A. (2012). *Native Listening. Language Experience and the Recognition of Spoken Words*. MA: MIT Press, Cambridge.

Di Cristo, A. (1998). Intonation in French. D. Hirst & A. Di Cristo (eds.), *Intonation Systems: A Survey of Twenty Languages*, CUP, pp. 195-218.

Di Cristo, A. (2000). Vers une modélisation de l'accentuation du français (seconde partie). *French Language Studies* 10, pp. 27-44.

Heiderscheidt, S. & P. Hiligsmann (2000). De accentuering in de tussentaal van Franstalige leerders van het Nederlands. *Leuvense Bijdragen,* 89(1/2), pp.17-131.

Lacheret-Dujour, A. & F. Beaugendre (1999). *La prosodie du français*. CNRS Editions, Paris.

Landis, J.R. & G.G. Koch (1977). The measurement of observer agreement for categorical data. *Biometrics* 33(1), pp. 159-174.

Leyden, K., van & V.J. van Heuven (1996). Lexical stress and spoken word recognition: Dutch versus English. *Linguistics in the Netherlands 1996*, pp.159-170.

Michaux, M.-C., Ph. Hiligsmann & L. Rasier (2012). Het klemtoonpatroon in de tussentaal van Franstalige leerders van het Nederlands [The stress pattern in the interlanguage of francophone learners of Dutch]. *XII. Internationaler Germanistenkongress* 2010, Warsaw, pp. 321-332.

Peperkamp, S. & E. Dupoux (2002). A typologocial stress 'deafness'. In C. Gussenhoven & N.

Warner, (Ed.), *Laboratory Phonology 7*, Mouton de Gruyter, Berlin, pp. 203-240.

Rasier, L. (2006), *Prosodie en vreemdetaal-verwerving. Accentdistributie in het Frans en Nederlands als vreemde taal.* Diss. UCLouvain.

Rietveld, A. C. M. & V.J. van Heuven (2009). *Algemene Fonetiek [General Phonetics]*. Bussum, Coutinho.

Schwab, S. & J. Llisterri (2012). Are French speakers able to learn to perceive lexical stress contrasts? *17th ICPhS,* City University of Hong Kong, pp. 1774-1777.

Trommelen, M. & W. Zonneveld (1989). *Klemtoon en metrische fonologie [Stress and metrical phonology]*. Muiderberg, Coutinho.

Evaluation of automatic prosodic segmentations

Klim Peshkov, Laurent Prévot, Roxane Bertrand

klim.peshkov@gmail.com, laurent.prevot@lpl-aix.fr, roxane.bertrand@lpl-aix.fr

Aix-Marseille Université and CNRS
Laboratoire Parole et Langage
5 Avenue Pasteur
Aix-en-Provence, France

Abstract

This paper presents a quantitative evaluation of existing automatic tools for prosodic segmentation of French speech. Two natural language processing evaluation metrics are proposed. It also compares performances of the tools on the whole corpus with the performances on narratives and low disfluency zones.

1. Introduction

The goal of this paper is to present an evaluation of existing tools for segmentation of French speech into prosodic units. More specifically, we test performances of several algorithms by comparing their outputs to an expert annotation of intonation phrases (IP). By performing evaluations on different data subsets, such as only narrative zones or zones with low disfluency rates, we would like to identify more precisely which phenomena make prosodic segmentation difficult. Moreover, in a previous prosodic breaks annotation campaign performed (Peshkov et al. 2012) we obtained only a fair inter-coder agreement. We suspect that it was largely due to the presence of disfluencies since the instructions for dealing with them were not very detailed. Thanks to our automatic identification of disfluencies we can verify this hypothesis.

Among the systems designed to extract information about French prosody, two are designed to detect prosodic units using acoustical features: Analor (Avanzi et al. 2008a) and the algorithm proposed in Degand and Simon (2009). Both systems are rule-based. However, little is known about their performances on different types of data. A quantitative evaluation of these tools on our conversational corpus should therefore yield interesting results.

It should be noted that the existing segmentation tools we considered (in particular Analor) seem to be designed for segmenting units larger than IP. Nevertheless, we consider that they constitute an interesting starting point and therefore we would like to assess their performance for IP in order to improve them and/or learn from their weaknesses on such a task.

The paper is organized as follows. Section 2 describes different automatic segmentations. Section 3 details the determination of low-disfluency zones based on morpho-syntactic labels. The results of the evaluations are presented in section 4.

2. Prosodic segmentation

2.1 Baseline

In order to understand the usefulness of the tools, we compare them to a baseline obtained corresponding to standard inter-pausal units (IPUs) with a silence threshold of 200 milliseconds.

2.2 DS algorithm

Following the work of Degand and Simon

(2009), we re-implemented their algorithm of prosodic segmentation (henceforth DS). The procedure described in the original proposition is semi-automatic. During the automatic phase, the algorithm loops through the syllables and assigns a boundary if one of the following conditions is satisfied: syllable lengthening (more than 3 times than the context mean), presence of a silent pause after the syllable (more than 200 milliseconds), intra-syllabic F_0 rise (more than 4 semi-tones) and syllable's mean pitch prominence (more than 5 semi-tones higher than the context mean). Some of the automatically detected boundaries are then manually suppressed if the syllable marks a hesitation or does not correspond to a final syllable of a lexical word.

We incorporated the prosodic part of these rules in a Python script designed to process the whole corpus at once. The script outputs segmentations in TextGrid format. The procedure is automatic. The pitch values are obtained using Prosogram's psychoacoustic F_0 stylisation (Mertens 2004). Our DS segmentation differs from the original in two aspects. First, since hesitations are difficult to detect automatically, the boundaries are suppressed at hesitations only if an *euh* (French hesitation marker) is present in the annotation. Second, we do not take into account lexical versus non-lexical word criterion.

2.3 Analor

Analor (Avanzi et al. 2008a) is a tool for detecting prosodic events in French speech corpora. Its primary functions are: (i) prosodic segmentation in periods, (ii) prominent syllables detection. In this paper we are concerned only in the first one.

We have previously run Analor on our data with standard parameters. The resulting segmentation gave rise to units, which were substantially longer than intonation phrases. For the present work, we were interested in tuning the different parameter to obtain shorter units. This is done in order to test how close Analor's segmentation can get to the intonation phrases and to learn the general behaviour of the tool on our data.

The segmentation algorithm of Analor has three main parameters : (i) pause length (in milliseconds); (ii) amplitude of F_0 movement, i.e. difference in height between the last F_0 extremum and the mean F_0 over the entire portion of the signal preceding the pause (in semitones); (iii) 'jump', difference between the last F_0 extremum preceding the pause and the first F_0 value following the pause (in semitones). Each cue is associated with three thresholds. This enables to capture the amount of cue activation. The decision to segment is taken after combining the weights of the three parameters. For example, if two of the parameters are greatly above the threshold, they can compensate for a low value of the third cue. For further details, see Avanzi et al (2008b). The standard values are shown in the first line of the table 1.

	Pause			Movement			Jump		
3	*250*	*330*	*660*	*3.0*	*5.0*	*8.0*	*2.0*	*3.5*	*7.0*
2	150	198	396	2.0	3.3	5.3	1.5	2.6	5.3
1	50	66	132	1.0	1.7	2.7	1.0	1.8	3.5

Table 1.Three settings for each of the three parameters

It was chosen to test 2 additional lower values for each parameter, which gave 3 levels for each parameter (see Table 1). All 27 possible combinations were tested. To produce new parameter sets, we chose the smallest value for each parameter that we wanted to test: 50 ms for pauses, 1 semitone for movement, 1 semitone for jump. The other values were calculated using the same relative distances as for standard values. In order not to overcomplicate the results presentation, we will show only three Analor outputs.

3. Determination of low disfluency zones

The disfluencies in the corpus were

detected by categorizing syntactic chunks (Peshkov et al 2013). The detection was based on a manually corrected automatic syntactic tagging (Rauzy and Blache 2009).

Chunking is an easy-to-implement and robust method for shallow syntactic analysis. Its main principle consists in including in one unit all the constituents situated to the left of each syntactic head (Abney 1991). As in our case we work with spoken material, silent pauses (starting from 200 milliseconds) were also used for segmenting. After chunking, every resulting syntactic pattern was characterized as being 'normal', 'incomplete' (without a syntactic head) or 'excessive'. An example of an incomplete pattern is a sequence of determiners, followed by a pause. A sequence of determiners, followed by a noun is an 'excessive' pattern.

The low disfluency zones here were defined as continuous stretches of normal chunks of a minimal length of 15 seconds.

4. Results

4.1 Speech material

For this study we used the Corpus of Interactional Data (CID) (Bertrand et al. 2008). It is made of 8 conversations of one hour involving two speakers. The protocol for obtaining this data was designed to enable highly natural interactions featuring a lot of overlaps and disfluencies. The speakers have very different speaking styles. The reference annotation is an annotation of intonation phrases (IP) performed by an expert linguist (Nesterenko et al. 2010).

Evaluations presented here have been performed on one dialogue of the corpus. Each speaker was recorded separately, which enables individual treatment. We used three datasets for the evaluation: (i) all data: the whole one hour dialogue; (ii) manually annotated narratives subset, consisting of 13 narratives, with total duration of 36 minutes; (iii) subset of low-disfluency zones (cf. section 3).

4.2 Evaluation measures

Precision and recall are conventional evaluation metrics from information retrieval. Precision corresponds to the percentage of correct boundaries in the evaluated segmentation. Recall is the percentage of boundaries in the reference segmentation detected by the algorithm.

$$\text{Precision} = \frac{\text{correctly detected boundaries}}{\text{number of boundaries in algorithm's segmentation}}$$

$$\text{Recall} = \frac{\text{correctly detected boundaries}}{\text{number of boundaries in reference segmentation}}$$

We are interested in comparing automatic outputs with the reference annotation done using a phonological approach to prosodic phrasing. Each phrase can be described in terms of metrical or acoustical cues, which can characterize either the beginning or the end of the phrase. This is why it seems interesting to have independent measures for the left and right borders of the units and the entire unit matches. The method involving these separate measures was introduced for evaluating the clausal unit detection in (Tjong et al. 2001). We adopted a delta of 160 milliseconds (average syllable length) to tolerate mismatches of approximately one syllable.

When used for segmentation evaluation, information retrieval metrics have a serious drawback. They do not take in consideration the distance between the borders of the segmentations being compared. Near-miss errors are penalized as heavily as insertion or deletion of borders and using delta can result in a bias. WindowDiff metrics was introduced to address this problem (Pevzner 2002). The algorithm operates as follows. It consists in moving a fixed-length window along the two segmentations, one unit at a time. For each position, the algorithm compares the numbers of borders in both segmentations. If the number of borders is not equal, the difference of the numbers is added to the evaluated algorithm's penalty. The sum of penalties is

then divided by the number of stops, yielding a score between 0 and 1. The score 0 means that the segmentations are identical.

WindowDiff was created for text segmentation tasks. When applying it to prosodic units evaluation in time-aligned transcripts, we had to adapt it to our case by introducing a time-based step. Results shown below were obtained with a step of 50 milliseconds.

4.3 Evaluation: all data

Figure 1 shows precision and recall scores of five segmentations for the starts of the units, their ends and for both borders. Figure 2 shows WindowDiff scores of the automatic segmentations for two speakers. In both figures, the performances on the entire corpus are shown by solid lines. Analor's results are represented by 3 parameter combinations (from more sensitive to less sensitive): 112, 223, 333. Each number represent the level of one of the parameters in order: pause, movement, jump. The values are found in table 1. Thus, 112 corresponds to *pause* [50, 66, 132] ms, *movement* [1.0, 1.7, 2.7] semitones and *jump* [1.5, 2.6, 5.3] semitones. 333 parameter combination corresponds to default settings which should yield the largest units.

It should be noted that none of the tools differ greatly from the baseline (IPU). The DS algorithm shows the highest recall, but its precision is lower by almost the same amount for the beginnings and for the ends of units. For the whole units, its precision is equal to the baseline, while the recall is 3.6% higher.

According to WindowDiff metrics, the DS algorithm is a little closer to the reference segmentation than the baseline. Analor based segmentations stay below the baseline.

Low scores may be explained by the fact that tools were designed to detect higher-level units than intonation phrases. All the automatic segmentations contain longer units than intonation phrases.

4.4 Evaluation of the subset of narratives

The evaluation of algorithms' performances on the sub-corpora of narratives and of low disfluency zones is done to determine which parts of the corpus are easier for the detection of prosodic events. Later, we will be able to compare this distribution with the distribution of low and high agreement zones in a semi-naïve annotation of prosodic-breaks. The narrative sub-corpus was created using a manual annotation of the narratives.

The hypothesis that the narratives are easier to segment than the interactional parts was not confirmed. Both metrics show that the similarity with the intonation phrases in the narrative sub-corpus is lower for all tools than in the whole corpus. This can be explained by the fact that interactive parts, missing in narrative sub-corpus, feature a lot of long pauses caused by changes of turn, and all algorithms use these pauses as one of the main cues.

4.5 Evaluation of low disfluency zones

Both metrics show higher similarity to the manual segmentation on low disfluency sub-corpus. It means that low disfluency zones are easier for prosodic events detection. This may also be seen as indirect confirmation of the effectiveness of disfluency detection procedure.

5. Conclusion and perspectives

In the future, we would like to use the annotations of prosodic breaks produced by semi-naive annotators for the evaluation (Peshkov et al. 2012).

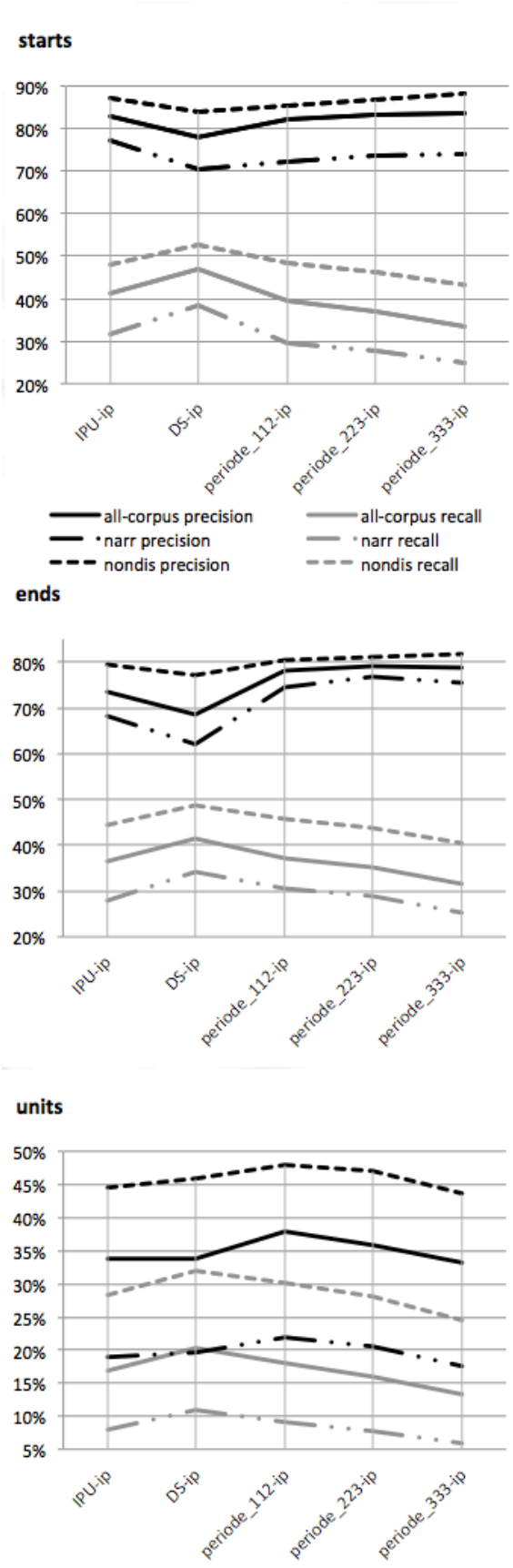


Figure 1 Precision and recall

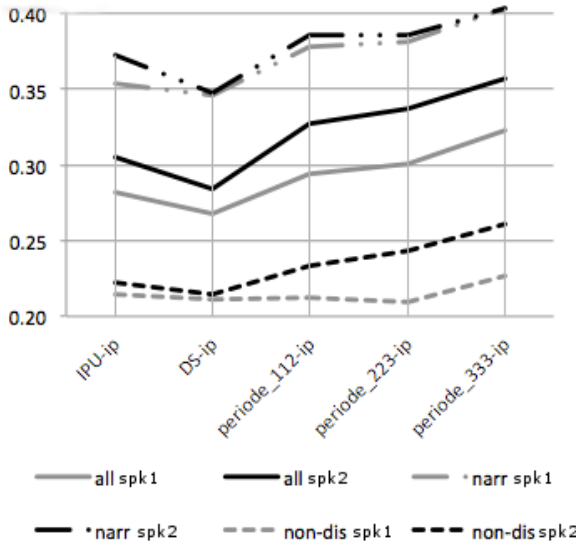


Figure 2 WindowDiff

A longer-term objective is to create an automatic segmentation system of spoken speech into discourse units. We plan to accomplish it by combining information from prosodic and syntactic levels.

As we would like to obtain even smaller units than as we did by tuning Analor's segmentation to be sensitive, it seems interesting to use Analor's prominence detection feature. Analor detects prominent syllables inside each period. Using them, one can divide periods into smaller units. We also want to explore the possibility to obtain prosodic segmentation using Momel-Intsint outputs (Hirst 2007).

Concerning evaluation measures themselves, we would like to compare the measures used here with the recent proposal of (Fournier and Inkpen, 2012).

Acknowledgments

This work has been realized with the support of the region Provences-Alpes-Côte d'Azur and of the ANR OTIM (Grant Number ANR-08-BLAN-0239). Thanks to Mathilde Guardiola, Anaïg Pénault and Irina Nestenrenko for annotation work and to Stéphane Rauzy for help with syntactic aspects.

References

Abney, S. (1991). Parsing by chunks. *Principle-based parsing*, vol. 44, pp. 257–278.

Avanzi, M. & Lacheret-Dujour, A. & Victorri, B. (2008a). Analor, a tool for semi-automatic annotation of French prosodic structure. *Proceedings of Speech Prosody'08*, pp. 119-122, Brazil.

Avanzi, M. & Lacheret A. & Victorri B. (2008b) Analor, un outil d'aide pour la modélisation de l'interface prosodie-grammaire. *Travaux linguistiques du CERLICO*, France, vol. 21, pp. 27-46.

Bertrand R. & Blache, P. & Espesser, R. & Ferr, G. & Meunier, C. & Priego-Valverde, B. & Rauzy, S. (2008). Le CID — Corpus of Interactional Data — Annotation et Exploitation Multimodale de Parole Conversationnelle. *Traitement Automatique des Langues,* 49 (3), pp. 1-30, France.

Degand, L. & Simon, A. C. (2009). On identifying basic discourse units in speech: theoretical and empirical issues. *Discours*, 4.

Fournier, C. & Inkpen, D. (2012). Segmentation Similarity and Agreement. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 152-161.

Hirst, D. (2007). A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation. *Proceedings of the XVIth International Conference of Phonetic Sciences*, pp. 1233-1236.

Mertens, P. (2004). The prosogram: Semi-automatic transcription of prosody based on a tonal perception model. Bel, B. & Marlien, I. (eds), *Proceedings of Speech prosody*, Paris.

Nesterenko, I. & Rauzy, S. & Bertrand, R. (2010). Prosody in a corpus of French spontaneous speech : Perception, annotation and prosody-syntax interaction. *Prodeedings of Speech Prosody 2010-Fifth International Conference*, Chicago.

Pevzner, L. & Hearst, M. A. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28:1, pp. 19-36.

Peshkov, K. & Prévot, L. & Bertrand, R. & Rauzy, S., Blache, P. (2012). Quantitative experiments on prosodic and discourse units in the Corpus of Interactional Data. *Proceedings of SemDial 2012 (SeineDial): The 16th Workshop on the Semantics and Pragmatics of Dialogue*, pp. 181-183.

Peshkov, K. & Prévot, L. & Rauzy, S. & Pallaud, B. (2013) Categorizing syntactic chunks for marking disfluent speech in French language, *Proceedings of DiSS 2013*, in press.

Rauzy, S. & Blache, P. (2009). Un point sur les outils du LPL pour l'analyse syntaxique du français. *Proceedings of ATALA 2009.*

Tjong, E.F. & Sang, K. & Djean, H. (2001). Introduction to the CoNLL-2001 shared task: clause identification, *Proceedings of the 2001 workshop on Computational Natural Language Learning* 7, pp. 127-132.

VtoV: a perceptual cue for rhythm identification

Massimo Pettorino, Marta Maffia, Elisa Pellegrino, Marilisa Vitale, Anna De Meo

{mpettorino, mmaffia, epellegrino, vitalem, ademeo}@unior.it
University of Naples L'Orientale, Italy

Abstract

Current metrics for the quantification of speech rhythm take into account parameters not easily detectable by listeners. To overcome this limit, in this study we propose a new model based on a parameter that account for listeners' ability to discriminate between different rhythmic patterns.

Starting from the results of a spectro-acoustic analysis conducted on singing, we found that Perceptual Centres align close to Vowel Onset Points (VOP). To test the perceptual relevance of interval between two consecutive VOPs, that we call VtoV intervals, we analyzed a multilingual corpus of TV news, advertisings and recited speech. The signal was segmented into vocalic/consonantal portions and into VtoV intervals. The standard deviation of all parameters was calculated. The results of the analysis show that VtoV is a crucial parameter both to classify languages on a rhythmic basis and to account for intra-linguistic speech style variations.

1. Introduction

There is ample evidence in literature that languages differ considerably in the way they produce rhythmical contrasts. According to the way isochrony is realized languages have been traditionally classified into three main groups: syllable-timed, stress-timed and mora-timed languages (see Pike 1945; Abercrombie 1967; Ladefoged 1975).

In the first group, syllable duration tends to remain relatively constant and unstressed syllables cannot be drastically reduced. In stress-timed languages, by contrast, stressed syllables recur at equal intervals and there is a substantial degree of syllable duration variability (see Dauer 1983). In the third group the mora, a sub-syllabic unit consisting in one short vowel and any preceding onset consonants, serves as a basic unit of rhythmical organization.

Nevertheless, over the years the attempts to find experimental evidence supporting the notion of acoustic isochrony have largely failed (for a review, see Bertinetto 1989; Kohler 2009) primarily because of the many factors that influence spoken communication, modifying its temporal organization.

Moreover, the attribution of rhythm-classes to particular languages requires very accurate syllable boundaries identification. However, syllable boundaries are hard to identify particularly when they occur within 1. a silent interval of a long stop consonant, 2. a cluster consisting of a nasal plus a voiced stop. In this case, the signal does not contain any discontinuity for the effect of a full or partial nasalization of the voiced-stop.

A further point to be clarified on the concept of isochrony is whether the syllable duration is considered from the view point of articulatory production or from that of perception. In fact the perceptual duration of a syllable does not necessarily correspond to the duration of the articulatory gesture. For instance, in a voiceless stop CV syllable, the articulatory duration is longer than the perceptual one, because the articulatory mechanics begins before the onset of the acoustic signal.

An attempt to overcome some of these methodological limits is represented, among others, by the work of Ramus et al. (1999). Starting from many experiments on the listeners' ability to discriminate between

languages with different rhythmic patterns, they proposed a new method to assign languages to the three different rhythmic groups. To the purpose, they calculated the proportion of vocalic intervals within the sentence and the standard deviation of consonantal intervals. The results of their study have indicated that syllable, stress and mora-timed languages differ from each other in the percentage of vocalic portion (%V) and in the standard deviation of the durations of consonantal intervals (ΔC).

Despite the numerous methodological advantages to this procedure, the %V/ΔC model does not seem to account for listener's ability to discriminate between languages according to their rhythmic features. Listeners are unlikely to manage to calculate, even roughly, the percentage of vocalic intervals and the standard deviation of consonantal clusters in real-time. Consequently, there should be another parameter, perceptually detectable, enabling listeners to distinguish a rhythmic pattern from another.

As rhythm is the regular succession of prominences in time (see Marotta 2011), such a parameter should be then linked to the recurrence of audible signal discontinuities.

In this regard, a sizeable body of research has demonstrated the existence of prominent instants in the speech signal that are perceptually more salient than others. These instants, called Perceptual Centres or P-Centres (see Morton et al. 1976) correspond to a particular point within the syllable that perceptually corresponds to its "moment of occurrence" (see Marcus 1981).

Additionally, sequences of P-Centres are thought to underlie the perception and production of rhythm in perceptually regular speech sequences. However, physical correlates of P-Centres have not been firmly established (see Villing 2003) and their exact location is a current matter of experimental verification (see Villing 2010 for reviews).

2. The study

2.1. First experiment

The existence of P-Centres is particularly evident in singing. In this case, the tempo of the music requires the singer to produce each syllable at precise time points. But how can a syllable, which corresponds to a time interval, be synchronous with an instant? There should exist within that interval a perceptually prominent point which allows for such synchronization.

In order to answer this question, we asked a professional singer to record an Italian song going in time with the beats of a metronome (92 bpm). The spectro-acoustic analysis of the corpus, carried out by means of Praat, has shown that all beats align with the vowel onsets, thus confirming some data present in the literature (see Tuller and Fowler 1980). 74% of beats occurs within 0.005 s from the vowel onset, while 26% shifts on average 0.034 s ($\sigma = 0.008$ s) (fig. 1).

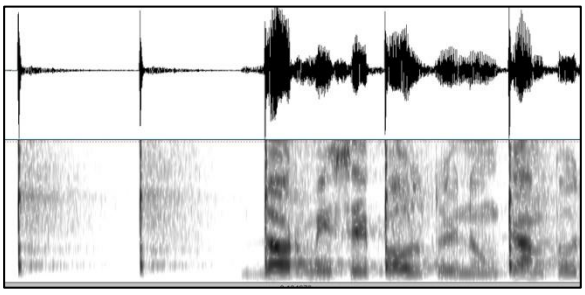


Figure 1

In order to collect further evidence that P-Centres tend to align close to the Vowel Onset Point (VOP), we examined spectro-acoustically some commercial songs played by very well-known artists. It is to be underlined that, unlike the song performed on the beats of a metronome, for songs with a complex instrumental accompaniment, it is not always possible to check on the spectrogram the synchronization between lyrics and music. However, where this synchronization was verifiable, the analyses have confirmed that P-Centres were located close to VOPs. It is therefore possible to

conclude that the VOPs represent those audible signal discontinuities that would guide listeners in the perception of rhythm. As a consequence, the interval between two consecutive vowel onset points (henceforth called VtoV interval) seems to be the cue enabling listeners to identify the rhythmic pattern of a language.

2.2. Second experiment

To test the role of VtoV in rhythm perception, we collected a multilingual corpus of about 15 minutes. The corpus was composed of TV news readings, speech taken from drug advertisings and recited speech. The languages were representative of the three rhythmic groups: Italian, French, English and Japanese.

As for the TV news, the speech samples were taken from RAI, RTF, BBC and NHK channels. Drug ad samples were drawn from the end of pharmaceutical television commercials when the voiceover recites the contraindications and side effects. Here, the speech is deliberately accelerated, sometimes through the use of signal manipulation, in order to hinder the full understanding of the message. As for the recited speech, the corpus consisted of verses from: 1) Shakespeare's 20th sonnet "A woman's face", 2) Montale's "Le quattro stagioni" and 3) Prévert's "Cet amour".

The entire corpus was segmented into vocalic/consonantal portions and into VtoV intervals on two separate tiers. The segmentation of glides followed the rules adopted by Ramus et al. (1999): [w] and [j] were treated as consonants and the boundary was placed between the approximant and the vowel. Falling diphthongs were segmented in one or two vowels intervals depending on the spectro-acoustic characteristics of the tract. If both vowels presented a specific steady-state formant pattern, the diphthong was divided into two VtoV intervals; otherwise, it was treated as a single interval.

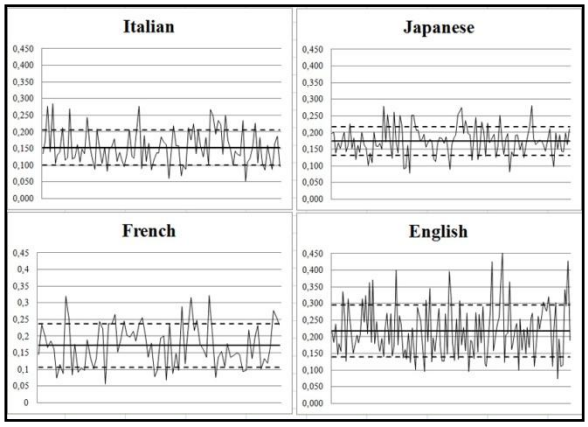


Figure 2

For each speech sample we calculated the average value of VtoV intervals, their standard deviation (ΔVtoV), the percentage of vocalic (%V) and consonantal (%C) portions and their standard deviations (ΔV and ΔC).

Figure 2 represents the sequence of VtoV intervals on *x* axis, and on *y* axis their duration (s). The data refer to TV news in Italian, French, Japanese and English. The continuous horizontal line indicates the average value of VtoV, while the dashed lines indicate the ΔVtoV. As the figure shows, English differs from the other languages, both for higher VtoV and for larger ΔVtoV. These results reflect the fact that English, compared to Italian, French and Japanese, is characterized by wider variety of syllable structure type, more complex consonant clusters, higher frequency of closed syllables and drastic reduction of unstressed vowels.

Tables 1 and 2 present VtoV and ΔVtoV for recited speech in Italian, French and English, in comparison with TV news. VtoV increases by about 40% in the three languages, and ΔVtoV undergoes an increase of 40% in English, of 14% in French, and of 55% in Italian.

	TV news (a)	**Recited speech (b)**	**Difference (b-a)**	**%**
Italian	0.153	0.213	+ 0.060	+39
French	0.172	0.250	+ 0.078	+45
English	0.215	0.299	+ 0.084	+39

Table 1

	TV news (a)	Recited speech (b)	Difference (b-a)	%
Italian	0.053	0.082	0.029	+55
French	0.065	0.074	0.009	+14
English	0.089	0.125	0.036	+40

Table 2

An opposite trend to recited speech was observed in the drug ads. The VtoV decreases by 38% for Italian and by 35% for French (Tab. 3). In both languages ΔVtoV, instead, decreases by 55% (Tab. 4).

	TV news (a)	Drug ad. (b)	Difference (b-a)	%
Italian	0.153	0.095	-0.058	-38
French	0.172	0.112	-0.060	-35

Table 3

	TV news (a)	Drug ad. (b)	Difference (b-a)	%
Italian	0.053	0.024	-0.029	-55
French	0.065	0.029	-0.036	-55

Table 4

Figure 3 shows the relationship between VtoV and ΔVtoV, analyzed *per* different languages and speech styles. Data indicate quite evidently that there is a direct relationship between the two variables: the higher the VtoV the larger ΔVtoV. Additionally, regardless of language and speech style, wider intervocalic intervals undergo greater variations. On the contrary the closer the vowels are, the more constant the intervocalic intervals.

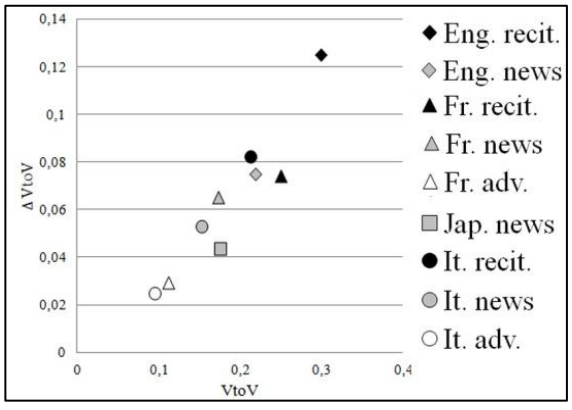


Figure 3

2.3. Third experiment

Another experiment was conducted to better investigate the relationship between VtoV variation and vocalic and consonantal variability. To the purpose, we plotted VtoV with ΔV and ΔC, and then ΔC with ΔV (fig. 4). From the three graphs it is possible to infer that VtoV variations are more greatly determined by ΔC rather than by the ΔV.

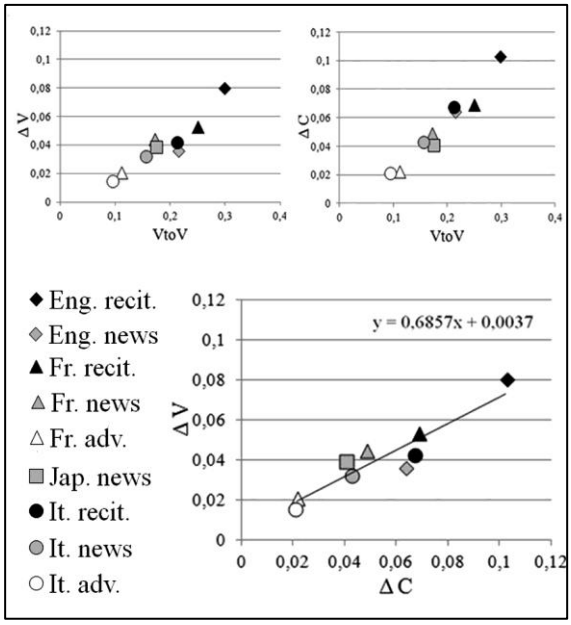


Figure 4

Given this direct relationship between VtoV and ΔC, we propose to revisit the %V/ΔC model by substituting ΔC with VtoV, the only perceptually salient and detectable parameter for listener. Therefore, we analyzed our multilingual corpus according to both models (%V/ ΔC and %V/ VtoV). The figures 5 and 6 show the results of both analyses. In the two graphs languages of different rhythmic groups are distributed along the *x* axis. From left to right there is English, stress-timed language, then Italian and French, syllable-timed languages, and then Japanese, isomoraic language. The distribution of languages along *y* axis indicates intra-language differences due to the diverse speech styles. In fact, going from the bottom to the top, the

rate of the speech samples moves from very fast to very slow.

To confirm whether differences in speech rate were recognized also on a perceptual level, 80 Italian listeners, aged between 18 and 23, were involved in a perception test. They were asked to evaluate the speech rate of 9 speech samples on a three-point scale (slow, medium and fast). To eliminate the message component, the samples were manipulated through lowpass filtering technique (cut-off frequency 400 Hz). The results show that, regardless of language and speech style, the excerpts judged as "slow" are those with a VtoV higher than 0.250 s; those recognized as "fast" corresponded to VtoV of about 0.1 s; those considered as "medium" were between 0.1s and 0.2 s.

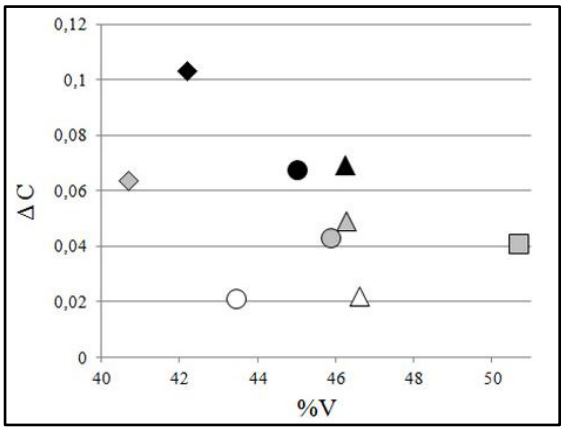


Figure 5

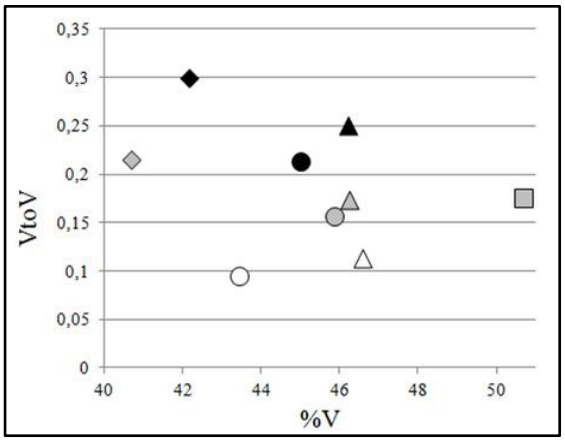


Figure 6

Nevertheless, it is important to underline that English presents VtoV values that are higher than those in French and Italian both for recited speech and news reading. This proves that VtoV is a parameter that does not only depend on speech rate but also on the rhythmic characteristics of the language. To test this hypothesis we will extend our analysis to other languages belonging to the three different rhythmic groups.

3. Conclusions

This study, performed on languages with different rhythmic characteristics and on different speech styles, shows that VtoV interval represents a relevant cue in the perception of rhythm. Under this perspective, the consonantal intervals can be considered as the interruptions or attenuations in the speech signal that determine those discontinuities underlying the perception of rhythm. These discontinuities, indeed, consist in the periodic recurrence of fully resonant vowel sounds. The %V/VtoV model is therefore very effective to represent the different rhythmic patterns of languages, providing a very articulated framework of the possible combinations among different languages and different types of speech. Data from our study seem to show that, speech style being equal, there is an inverse relationship between the two parameters: the higher %V, the lower VtoV.

In further steps of our research we will investigate whether the perception of speech rate, that is proved to be linked to VtoV variations, depends on the different rhythmic groups of languages. To the purpose, we will administer perception tests based on natural speech to native speakers of the target languages.

References

Abercrombie, D. (1967). *Elements of general phonetics*. Aldine, Chicago.

Bertinetto, P. M. (1989). Reflections on the dichotomy «stress» vs «syllable timing». Révue de Phonétique Appliquée 91, pp. 99-129.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, 11, pp. 51-62.

Kohler, K. J. (2009). Rhythm in speech and language. A new research paradigm. *Phonetica* 66, pp. 29-45.

Ladefoged, P. (1975). *A course in phonetics*. Harcourt Brace Jovanovich, New York.

Marotta, G. (2011). Ritmo, voce dell'*Enciclopedia dell'Italiano Treccani*, vol. II, Istituto dell'Enciclopedia Italiana, Simone Raffaele, 1262, 2011.

Marcus, S. M. (1981). Acoustic determinants of Perceptual-centre (P-Centre). *Perception and Psychophysics*, 30, pp. 247-256.

Morton, J., S. Marcus, & C. Frankish (1976). Perceptual Centers (P-centers). *Psychological Review*, 83:5, pp. 405-8.

Pike, K. L. (1945). *The intonation of American English*. Ann Arbor, Michigan: University of Michigan Press.

Ramus, F., M. Nespor, & J. Mehler (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, 72, pp. 1-28.

Tuller, B. & C. A. Fowler (1980). Some articulatory correlates of perceptual isochrony. *Perception and Psychophysics*, 27: 4, pp. 277-283.

Villing, R., T. Wardc & J. Timoney (2003). P-Centre Extraction from Speech: the need for a more reliable measure *Proceedings of ISSC 2003*, Limerick.

Villing, R. (2010) Hearing the Moment: Measures and Models of the Perceptual Centre http://eprints.nuim.ie/2284/1/Villing_2010_-_PhD_Thesis.pdf

Variation prosodique situationnelle : étude sur corpus de huit phonogenres en français

Tea Pršir[✈✲], Jean-Philippe Goldman[✈], Antoine Auchlin[✈]

tea.prsir@unige.ch, jean-philippe.goldman@unige.ch, antoine.auchlin@unige.ch
[✈]Département de Linguistique, Université de Genève
[✲]Institut Langage & Communication, UCLouvain

Abstract

This communication presents partial results from an on-going study on prosodic and phonostylistic variation across phonogenres (i.e., speaking styles), typified acoustic images associated to language activity-types. It extends previous work by Goldman et al. 2011, Simon et al. 2010, with a larger corpus (633 min.), a greater number and complementary repertoire of considered genres. It further joins rhythmical comparative measurements (Dellwo 2010) to Goldman et al.'s (2007) ProsoReport.

Lucci (1983) and Koch & Oesterreicher's (2001) situational dimensions set is reduced to situational conditions and features sufficient to distinguish considered phonogenres one from another.

Corpus treatment, annotation and measurement, is achieved semi-automatically, through a set of Praat implemented tools, and manual steps.

The communication presents corpus collection, annotation, treatment methodology, and results for studied phonogenres.

1. Introduction

Les situations dans lesquelles on parle laissent des marques dans la parole. Inversement, une partie des propriétés prosodiques indicielles (au sens sémiotique peircien de symptôme) est déterminée par les conditions de production de la parole. L'étude de la variation situationnelle de la parole a entre autres pour but d'établir des corrélations entre propriétés des situations et traits prosodiques. Si les situations se regroupent selon une typologie implicite, dont il s'agit de dessiner les contours, inversement, des traits prosodiques typiques contribuent à stabiliser des genres de parole (le reportage sportif ; le sermon ; etc.).

Cette contribution s'inscrit dans le cadre large de l'intérêt accru en sciences du langage pour la question des genres (Beacco 2004 ; Solin 2011) et dans le cadre restreint de l'étude des genres oraux, ou phonogenres, et de la variation situationnelle (Simon et al. 2010 ; Goldman et al. 2011 ; Boula de Mareüil 2012 ; Obin et al. 2008 ; etc.). Avec la même méthodologie semi-automatique, elle élargit le répertoire des genres considérés; elle se base sur un corpus plus substantiel que ceux sur lesquels se basent les études préalables; enfin, elle élargit l'inventaire des traits macro-prosodiques classiques à différents paramètres microprosodiques, selon les propositions de Dellwo (2010).

2. Genre et phonogenre

Nous nommons, à l'instar de Goldman et al. (2011), *phonogenre* une image acoustique associée à un certain type de situation et d'activité de parole, et *phonostyle* les propriétés d'un échantillon. Rassembler nos échantillons selon leurs situations de production et non leurs ressemblances phonostylistiques évite le cercle vicieux pointé par Beacco (2004 :111). Nous analysons et distinguons les situations en *traits situationnels*, inspirés des *invariants situationnels* (Lucci 1983) et des *traits conceptionnels* (Koch & Oesterreicher 2001), dans quatre dimensions[1].

[1] i. Caractère *médiatique* (exclusivement média, semi-, ou non média) ; ii. degré de *préparation* (lu, semi-préparé, spontané) ; iii. type d'*audience* (public, face-à-face, pas d'audience, micro) ; iv. degré d'*interactivité* (nulle, semi, ou interactif). Notre

3. Corpus, collecte, annotation

Suite à la constitution du corpus C-Prom (Avanzi et al. 2010), il s'est avéré nécessaire de contraindre davantage les situations de parole (pour éviter la dispersion) ainsi que de rassembler plus de locuteurs par phonogenre (pour éliminer l'idiosyncrasie). Ainsi, le corpus PhonoGenre est constitué de situations de parole plus circonscrites, et comprenant au moins 10 locuteurs par situation d'énonciation.

3.1. Constitution du corpus

Le corpus est constitué de 8 phonogenres : discours parlementaire [ASS] (questions au gouvernement à l'Assemblée Nationale française), conversation [CNV], didactique [DID], lecture [LEC], liturgique [LIT], revue de presse radiophonique [RPR], commentaire sportif [SPO], et vœux présidentiels du Nouvel An [VXP]. On pourrait considérer les genres ASS et VXP comme appartenant à un 'macro-genre' discours politique. Le corpus VXP contient en plus une dimension diachronique (de De Gaulle 1968 à Sarkozy 2007 pour les présidents français; et de Dreifuss 1999 à Calmy-Rey 2011 pour les présidentes suisses).

PhonoGenre	Nb extr.	Durée(mn.)
ASS	10	20
CNV	40	140
DID	14	84
LEC	40	112
LIT	7	54
RPR	15	93
SPO	5	35
VXP	15	95
TOTAL	146	633

Tableau 1 Nombre d'extraits et durée par phonogenre

Au final la durée moyenne des enregistrements est de 4mn20s (min. 2mn, max. 13mn). 75% d'entre eux sont d'une durée inférieure à 5mn. Le corpus est tiré de sources médiatiques variées : la télévision (DID, VXP, SPO), la radio (DID, RPR) ou l'internet (LIT, ASS) ; CNV et LEC proviennent du corpus C_PROM-PFC (Avanzi et al. 2012).

3.2. Alignement et annotation

L'intégralité du corpus est alignée au niveau des phonèmes, syllabes et mots à l'aide d'EasyAlign (Goldman 2011). Les variations stylistiques (liaison, élision, emploi du schwa, hésitation), les pauses avec et sans prise de souffle, ou les différents bruits de bouche ou extérieurs sont annotées dans une tire nommée *delivery*. Ces étapes préparatoires sont à la base des analyses acoustiques et statistiques.

4. Analyse acoustique: méthodologie, rapport prosodique

Cette première étude sur l'ensemble du corpus vise à rendre compte des caractéristiques prosodiques globales par phonogenre. Deux outils ont été associés pour produire 128 mesures acoustiques sur les 146 extraits sélectionnés. L'outil ProsoReport (Goldman et al. 2007) propose un rapport prosodique complet sur les deux principaux paramètres acoustiques de la prosodie (mesures tonales et temporelles) pour différentes unités de parole de taille variée telles que les segments phonétiques, les syllabes, les pauses et les unités séparées par les pauses (USP ou suites sonores), ainsi que pour l'enregistrement tout entier (taux d'articulation, durée moyenne et déviation standard des segments, syllabes, USP, distribution du registre, entre autres). De plus, une détection automatique des syllabes proéminentes permet d'obtenir une série de mesures complémentaires (proportion, écart intonatif et temporel des syllabes proéminentes). Parmi les 64 mesures acoustiques, seulement 51 sont pertinentes

analyse ici n'est pas centrée sur la part spécifique de chacun à travers les phonogenres – mais celle-ci fait partie de notre étude générale.

dans notre cas car non seulement elles comportent des mesures rationalisées (moyenne, taux et pourcentage) et non des mesures brutes (durée totale, nombre de syllabes, etc.), mais aussi elles permettent de comparer des groupes d'enregistrements contenant des locuteurs différents en ignorant des mesures propres à un locuteur comme le registre. Le second outil, décrit dans Dellwo (2010), s'oriente exclusivement sur des mesures temporelles et de variabilité rythmique, sur la base de la durée des intervalles vocaliques, consonantique et syllabiques. Pour ces variables, 51 mesures sont prises en compte sur les 58 fournies par l'outil. Ainsi nos données se résument en une table de 102 mesures acoustiques décrivant les 146 extraits.

5. Résultats - observations

Certains des 102 descripteurs prosodiques calculés sont assurément plus pertinents que d'autres. Nous les avons classés en opposant leur domaine prosodique (macro- ou micro-) et le paramètre acoustique associé (tonal ou temporel). Certaines mesures sont décrites en détail. Puis nous reprenons l'ensemble des descripteurs prosodiques pour une analyse en composantes principales (Fig. 4).

5.1. Mesures macroprosodiques

Il existe un effet global du taux d'articulation en syll/sec, ($F(7,138) = 7.453$, $p<0.001$). Les tests post-hoc opposent significativement un groupe VXP et LIT aux autres (Fig. 1).

D'autres mesures macroprosodiques, notamment mélodiques, font ressortir des proximités entre les phonogenres. La déviation standard de f0 est plus élevée pour DID, RPR et VXP, ce qui signifie une agitation mélodique plus importante. Au contraire, ASS, CNV et LEC sont beaucoup moins sujets aux variations mélodiques. Quant au registre tonal, l'étendue (Q05-Q95) est plus importante pour DID, RPR et VXP, indice d'une expressivité prosodique plus importante que pour ASS et LEC.

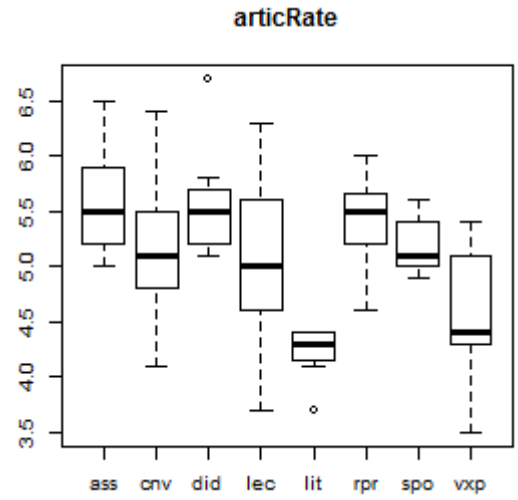


Figure 1 Distribution du taux d'articulation (en syll/sec) pour les 8 genres

La mesure du nombre de syllabes par suite sonore (Fig. 2) présente un regroupement différent: ASS, DID et RPR se distinguent par le nombre plus élevé et sont en contraste avec LIT, SPO et VXP. Effectivement d'autres mesures confirment la tendance des trois premiers genres à produire des suites sonores plus longues que les trois derniers.

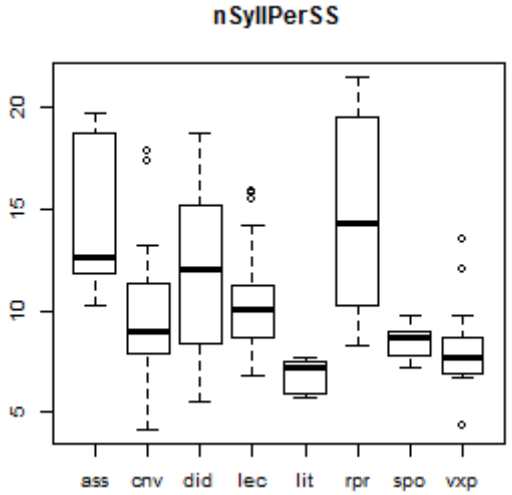


Figure 2 Distribution du nombre de syllabes par suite sonore (en syllabes) pour les 8 genres

5.2. Mesures microprosodiques

DID, LEC, RPR et SPO ont une proportion de syllabes montantes significativement plus importante ($F(7,138) = 14.53$, $p<0.001$) que

celle de CNV, LIT et VXP. Nous supposons que le taux faible pour CNV, LIT et VXP est lié à la dimension 'empathique' de ces genres. Quant aux syllabes descendantes, DID, LIT et VXP ont un pourcentage plus élevé, sans qu'il y ait un grand écart par rapport aux autres genres.

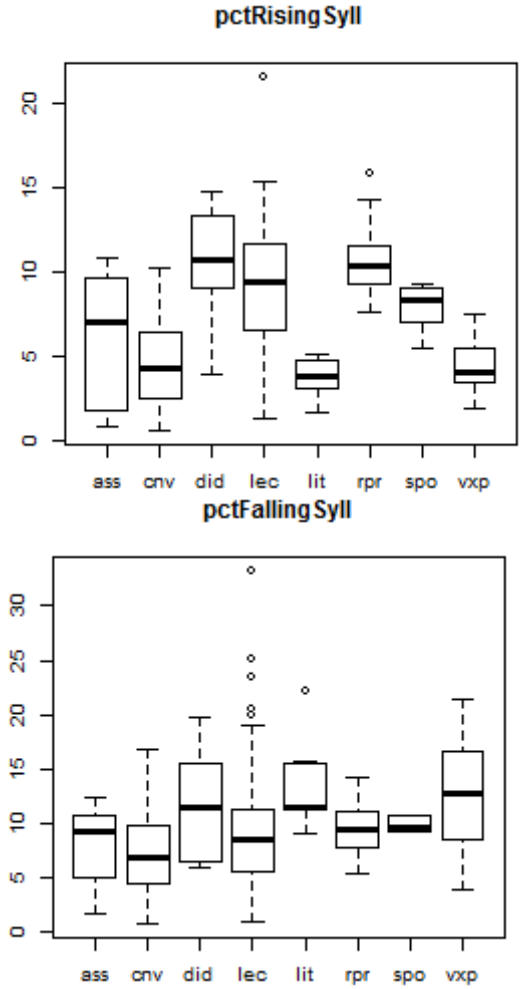


Figure 3a et 3b Proportion (en %) de syllabes montantes (3a) et descendantes (3b), 8 genres

Notons aussi une corrélation entre durée moyenne des syllabes et leur déviation standard, qui isole ASS et RPR (dur. moy. 0.18 s. et dév. de 0.08), reflet de leur caractère temporellement contraint. À l'opposé, VXP et LIT (dur. moy. 0.23 s., dév. entre 0.11 et 0.12) sont plus libres.

Une analyse en composantes principales (CP) a été réalisée sur l'ensemble du corpus. Les deux premières CP expliquent 58% de la variance, alors que les dix premières en expliquent 90.5%. Une analyse discriminante dans un but de classification automatique montre que 95% des extraits sont correctement identifiés.

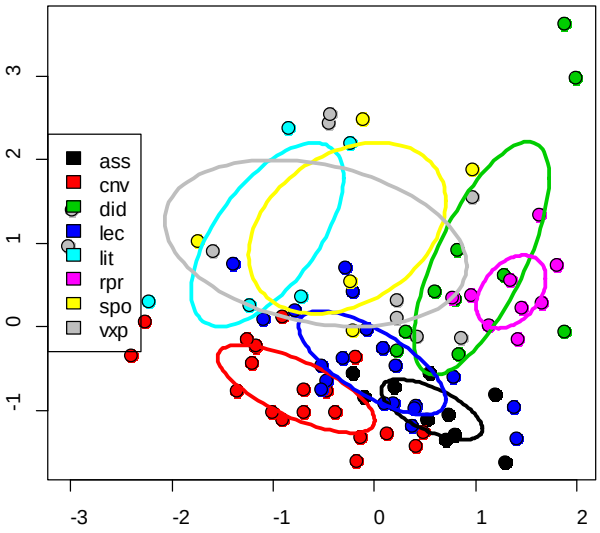


Figure 4 Tracé des 89 extraits dans les 2 premières composantes principales et les ellipses correspondant à un intervalle de confiance de 0.6

Le graphique en nuage de points (Fig. 4) représente la projection à une sélection réduite à 89 extraits selon les 2 premières composantes principales. Cette réduction du corpus initial a consisté à éliminer les extraits multiples avec le même locuteur afin de rendre compte de la dispersion phonogenre et non idiostyle. Remarquons que ASS (en noir) est le phonogenre le plus compact. La dispersion de LIT (en bleu) correspond également à deux situations du discours liturgique : celle de la messe (en haut du graphique) et celle de la lecture de l'homélie enregistrée, diffusée sur internet. RPR (en rose) est plutôt compact, mis à part un locuteur dont les trois extraits se trouvent isolés. Il s'agit d'un journaliste avec un discours particulièrement expressif. SPO est le phonogenre le moins compact. Les trois sports représentés ici (foot, rugby et basket) possèdent des dynamiques cinétiques différentes qui sont reflétées dans la parole (Audrit et al. 2012). Enfin, VXP est un phonogenre qui manifeste plusieurs particularités : d'un côté (gauche, haut) sont

regroupés quatre président-e-s suisses hommes et femmes confondus ; à l'extrémité droite se trouvent deux présidents français avec un phonostyle qui peut être dû à l'ancienneté des enregistrements (Pompidou 1969 et Giscard d'Estaing 1975) ; Sarkozy se trouve isolé du reste des présidents regroupés au milieu du graphique.

6. Discussion

Plusieurs mesures rapprochent les caractéristiques prosodiques de LIT et de VXP. Une analyse discursive confirme leur rapprochement aussi au niveau du contenu. Nous avons vu aussi que VXP et ASS s'opposent par plusieurs paramètres malgré leur appartenance commune au macrogenre politique. De manière générale, les résultats ont mis en évidence une proximité entre DID, RPR et VXP pour la plupart des caractéristiques prosodiques, les opposant au groupe ASS, CNV et LEC. Une explication envisagée serait le caractère médiatique des premiers.

Les mesures présentées ici montrent l'importance, dans la définition du phonogenre, des traits situationnels, que nous étudierons dans une prochaine étude. Enfin, la notion non abordée d'émaillage (c'est-à-dire la manifestation non continue, mais ponctuelle, de caractéristiques phonostylistiques) nous incite également à orienter les recherches vers une caractérisation dynamique des échantillons de parole.

Références

Avanzi, M., A.C. Simon, J.-P. Goldman & A. Auchlin (2010). C-PROM. Un corpus de français parlé annoté pour l'étude des proéminences, *Actes des 23èmes journées d'étude sur la parole* (Mons, Belgique, 25-28 mai 2010).

Avanzi, M., S. Schwab, P. Dubosson & J.-P. Goldman (2012). La prosodie de quelques variétés de français parlées en Suisse romande. Simon, A.C. (ed.). *La variation prosodique régionale en français*. Bruxelles, De Boeck/Duculot, pp. 89-119.

Audrit, S., T. Pršir, A. Auchlin, & J.-P. Goldman (2012), Sport in the media: a contrasted study of three sport live media reports with semi-automatic Tools, *Speech Prosody* 2012.

Beacco, J.-C. (2004). Trois perspectives linguistiques sur la notion de genre discursif. *Langages* 38/153, pp. 109-119.

Boula de Mareüil, Ph. (2012). Accents et styles. Une étude à base de perception et d'analyses acoustiques à travers le traitement automatique de la parole. HDR, Université Paris 3.

Dellwo, V. (2010). Influences of speech rate on the acoustic correlates of speech rhythm: An experimental phonetic study based on acoustic and perceptual evidence. PhD-Dissertation, Universität Bonn.

Goldman, J.-P. (2011) EasyAlign: an automatic phonetic alignment tool under Praat. *InterSpeech* September 2011, Florence, Italy.

Goldman, J.-P., A.C. Simon, A. Auchlin & M. Avanzi (2007). Phonostylographe, un outil de description des phonostyles prosodiques. *NCLF* 28, pp. 219-237.

Goldman, J.-P., A. Auchlin & A.C. Simon (2011). Discrimination de styles de parole par analyse prosodique semi-automatique. Yoo, H.-Y. & E. Delais-Roussarie (eds.) *Actes d'IDP 2009*. Paris, Septembre 2009.

Koch, P. & W. Oesterreicher (2001). Langage parlé et langage écrit. Holtus G., M. Metzeltin & Ch. Schmitt (eds.), *Lexikon der Romanistischen Linguistik*, I/2. Niemeyer, Tübingen, pp. 584-627.

Lucci, V. (1983). Étude phonétique du français contemporain à travers la variation situationnelle. Université des langues et lettres, Grenoble.

Obin, N., A. Lacheret-Dujour, C. Veaux, X. Rodet & A.C. Simon (2008). A Method for Automatic and Dynamic Estimation of Discourse Genre Typology with Prosodic Features. *InterSpeech Proceedings*, pp. 1204-1207.

Solin, A. (2011). Genre. Zienkowski, J., J.-A. Ostman & J. Verschueren (eds.), *Discursive Pragmatics*. John Benjamins, Amsterdam, pp. 119-134.

Simon, A.C., A. Auchlin, M. Avanzi & J.-P. Goldman (2010). Les phonostyles: une description prosodique des styles de parole en français. Abecassis, M. & G. Ledegen (eds.), *Les voix des Français. En parlant, en écrivant*. Peter Lang, Berne, pp. 71-88.

Étude prosodique des périodes au sein d'une tâche de narration d'histoires imagées en séquence

Lucie Rousier-Vercruyssen, Anne Lacheret-Dujour, Marion Fossard

lucie.rousier-vercruyssen@unine.ch[1], anne@lacheret.com[2], marion.fossard@unine.ch[1]
[1]Université de Neuchâtel, Suisse
[2]Université de Paris Ouest Nanterre la Défense, France

Abstract

The influence of prosody on the segmentation of utterances is often emphasized in the literature. In this paper, we study the periodic segmentation in a storytelling in sequence task performed by 12 native French speakers from Switzerland. Our aim is to explore the link between the segmentation of a discourse into major prosodic unity called "periods" and reference marking in discourse.

1. Introduction

En français parlé, parmi les travaux consacrés aux indices prosodiques et syntaxiques dans la segmentation du continuum sonore en unités discursives élémentaires, ceux qui portent sur le marquage de la structure informationnelle sont extrêmement féconds. La question des relations entre constructions prosodiques et expression de la référence dans les textes reste, en revanche, encore peu décrite. C'est cette thématique qui fait l'objet de notre communication. Plus spécifiquement, notre étude vise à analyser le marquage linguistique de la référence, i.e. les corrélations entre le statut ± accessible d'un référent et les traces syntaxiques et prosodiques qu'il laisse dans le message, lors d'une tâche d'élaboration narrative produite en situation d'interaction. Du point de vue syntaxique, il s'agit d'étudier la structure interne des expressions référentielles, i.e. les unités morphosyntaxiques qui les constituent, en suivant l'hypothèse que cette structure interne est corrélée au degré d'accessibilité des référents (Ariel 1990). Sur le versant prosodique, si les angles d'attaque sont a priori pluriels, nous ne nous perdrons pas dans la complexité des observables acoustiques pour cette étude préliminaire, et nous restreindrons volontairement notre analyse à la segmentation des énoncés en unités prosodiques majeures que nous appelons *périodes intonatives*. Nous proposons donc une approche globale, focalisée sur l'organisation du flux de parole dans le temps. Autrement dit, il s'agit de voir dans quelle mesure l'expression de la référence conditionne la segmentation du discours en unités prosodiques élémentaires et l'organisation interne de ces dernières (Chafe 1998, Smith et al. 2005).

Pour une tâche qui a pour but de raconter une histoire à partir d'une succession d'images, chaque image décrivant une action accomplie par des personnages, nous formulons l'hypothèse que la segmentation en périodes est déterminée à la fois par la nature et la distribution des expressions référentielles dans le flux de parole, et par la façon dont un locuteur décrit le passage d'une action à une autre. Du point de vue intonosyntaxique, se pose la question du poids respectif de l'un ou l'autre de ces deux paramètres dans la segmentation en périodes. Si le but communicatif premier est de décrire la succession des actions, il est probable que la segmentation soit guidée par le découpage en unités de rection (noyau verbal, éléments régis et satellites éventuels) ; les périodes intonatives s'aligneraient donc sur les unités syntaxiques (au moins une unité de rection dans une période). S'il s'agit plutôt de structurer le discours autour des entités référentielles et selon leur saillance relative, i.e. leur statut

± actif[1], on peut s'attendre à des phénomènes de restructuration, ou fragmentation syntaxique, les différentes unités qui s'organisent autour du noyau rectionnel pouvant former des périodes autonomes. Reste à savoir, toute chose égale par ailleurs, dans quelle mesure il n'existe pas un ou des principes qui paramètrent la segmentation des périodes en fonction du poids, lourd vs léger, des expressions référentielles qui se succèdent dans la chaîne parlée. Pour préciser ce point, plus un marqueur référentiel sollicite de matériel de codage syntaxique, plus il est lourd et, à l'inverse, moins il en réclame, plus il est léger[2]. D'autre part, en suivant les principes d'Ariel (1990), un marqueur lourd a la caractéristique d'être rigide, très spécifique, très informatif, non atténué ; un marqueur léger est atténué, peu spécifique, peu informatif. Sur ces bases, se pose la question de savoir si plusieurs marqueurs rigides et très informatifs peuvent exister dans une même période. Autrement dit, la segmentation en périodes n'est-elle pas pilotée par un principe d'alternance des différents types de marqueurs ?

2. Le corpus

Le corpus est extrait d'un projet financé par le Fond National Suisse (FNS), nommé « Discours et théorie de l'esprit : utilisation d'indices référentiels et prosodiques pour évaluer l'attribution de connaissances aux autres en situation d'interaction verbale. » et dirigé par Marion FOSSARD (Neuchâtel) (http://p3.snf.ch/project-140269). Il est composé de deux histoires extraites d'un ensemble, présentées sous forme de 6 images en séquence mettant en scène deux personnages dont la saillance est manipulée au cours de l'histoire (personnage en focus ou en arrière-plan). Différemment des tâches plus classiques de narration (ex : « Frog where are you ?», Bamberg, 1987) dans lesquelles le locuteur raconte une histoire à un interlocuteur, sans but collaboratif, la tâche de narration d'histoires imagées en séquence est utilisée avec le paradigme de communication référentielle (Champagne-Lavau et al. 2009; Clark et al. 1986). Ce paradigme permet de recréer une situation d'interaction verbale collaborative entre deux partenaires séparés par un écran opaque, un locuteur et un interlocuteur. Le but du locuteur est de permettre à son interlocuteur d'ordonner les items, classiquement des tangrams[3], dans le même ordre que le sien. L'utilisation de tangrams présente toutefois certaines limites comme une production peu diversifiée des marqueurs nominaux et une description d'images plutôt qu'une narration. Liant le paradigme de communication référentielle et les histoires en séquence, la tâche de narration d'histoires en séquence permet d'évaluer la façon dont un participant planifie son discours et, notamment, de déterminer le type d'information discriminante qu'il produit pour permettre à un destinataire d'identifier et d'ordonner les 6 images qui constituent la séquence de l'histoire. Le matériel ayant été spécifiquement développé pour manipuler la saillance des personnages, nous n'étudierons que la référence à ceux-ci. La durée totale du corpus d'étude est de 20 minutes représentant 2830 mots, soit 3039 syllabes et 183 périodes.

2.3. Les participants

Douze participants suisses (moyenne d'âge : 49,75 ; étendue 19-83 ans), 6 hommes et 6 femmes, de langue maternelle française (de Suisse romande), ont participé à cette expérience.

[1] Voir chez Chafe, (1974), le principe de « spotlight of consciousness ».

[2] Par exemple, un groupe nominal étendu, lourd, s'oppose à un pronom atone, léger.

[3] Formes géométriques non définissables à priori.

2.4. Analyse des données et annotations

✓Le codage des marqueurs référentiels

Plusieurs études (Ariel 1990, 2010; Gundel et al. 2012; Cornish 2010) ont montré que l'utilisation d'un marqueur référentiel particulier dépend du statut de l'information (connue/nouvelle) et de son activation (en focus ou non) au sein du discours. Suivant l'échelle d'accessibilité d'Ariel (1990) en particulier, lorsqu'une information est nouvelle et peu accessible, le marqueur attendu serait un marqueur lourd (un groupe nominal et ses modifieurs par exemple). Si l'information est connue et très accessible au sein du discours (dans le focus d'attention), il est attendu que ce soit un marqueur léger (un pronom, par exemple) qui code ce statut. Une étude (Gonzalez et al. 2011) utilisant le même matériel que le nôtre a confirmé cette hypothèse. Dans la phase d'introduction des personnages (information nouvelle), les marqueurs lourds (indéfinis principalement) étaient plus utilisés que lors des phases de maintien/changement de ces personnages (en focus ou non). A l'inverse, les marqueurs légers (pronoms et anaphores zéros) étaient davantage produits lors des phases de maintien, alors que dans les phases de changement (mise en focus d'un personnage d'arrière-plan), des marqueurs lourds (définis principalement), étaient les plus utilisés. Il existerait donc un lien entre les marqueurs référentiels et les différentes étapes de la construction du discours. Ce phénomène serait lié au savoir partagé co-construit par les locuteurs (Colle et al., 2007).

Pour l'annotation référentielle[4] de notre corpus, les marqueurs sont codés de la façon suivante, du plus lourd au plus léger :

- 'in', les marqueurs d'accessibilité basse : syntagme nominal indéfini;
- 'd+', les marqueurs d'accessibilité intermédiaire : syntagme nominal défini; syntagme nominal démonstratif; syntagme nominal possessif; pronom démonstratif; pronom possessif; pronom disjoint;
- 'cz', les marqueurs d'accessibilité élevée : pronom personnel anaphorique; pronom relatif; anaphore zéro.

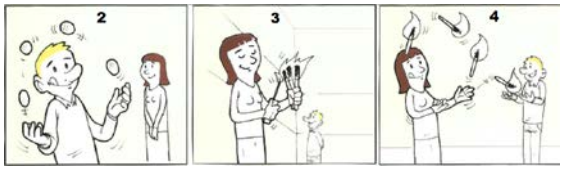


Figure 1 Exemple d'images (2 : phase de maintien, 3 : phase de changement, 4 : phase de maintien) d'une histoire de niveau 2

✓La segmentation en périodes

La segmentation en périodes est réalisée avec l'outil ANALOR (Lacheret et Victorri 2002, Avanzi et al. 2008), sur les bases de critères acoustiques stables et indépendamment de toute détermination syntaxique préétablie. En pratique, l'algorithme prend en compte les variations mélodiques globales et locales dans un intervalle de temps donné, soit les trois paramètres suivants : (i) la présence d'une pause silencieuse d'une certaine durée, (ii) l'amplitude du geste mélodique qui précède la pause, mesurée par l'intervalle en demi-tons entre le dernier extremum de F0 et sa moyenne sur toute la portion qui précède la pause, (iii) l'amplitude du saut définie comme la différence de hauteur mélodique entre le dernier extremum de F0 précédant la pause et la première valeur de F0 suivant la pause. A chaque paramètre est associé un seuil de coupure évalué selon une échelle de valeur entre -1 (coupure peu souhaitable) et + 2 (coupure très souhaitable). A partir de là, l'algorithme permet une certaine souplesse dans la prise de décision. En effet, celle-ci ne dépend pas de la valeur précise des seuils, mais uniquement de leur ordre de grandeur : quand l'un des paramètres est très proche du seuil (qu'il soit au-dessus ou au-dessous), la

[4] Équivalente à celle produite par Gonzalez, S. et al. (2011).

décision de coupure est prise en fonction de la situation d'ensemble.

Deux types d'opérations émergent des annotations: la première conduit à produire des périodes qui contiennent au moins une unité de rection *(il jongle)* ; la seconde fragmente les unités rectionnelles en plusieurs périodes *(sa <femme> <lui lance alors le défi>)*[5]. Nous parlerons donc respectivement de périodes alignées avec la syntaxe et de périodes fragmentées.

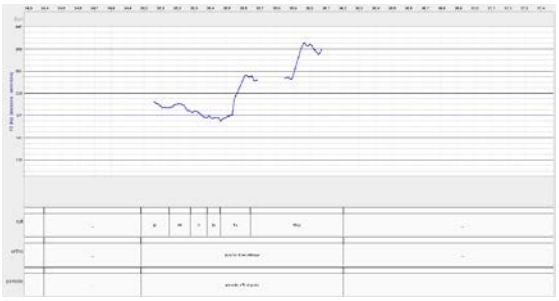


Figure 2. Visualisation sous Analor de la période 'puis lui il les rattrape'. Avec de haut en bas les variations de la fréquence fondamentale et les tires d'annotation : (tire syllabique, tire orthographique, tire période)

3. Action ou référence privilégiée ?

Les périodes alignées représentent 74,32% (136/183*100) de la totalité des périodes étudiées dans cette étude. Les périodes fragmentées : 25,68 % (47/183*100). La différence entre les deux types de construction est significative (ANOVA univariée, $F(1,7)=29.964$, $p<0.05$).On peut donc en conclure que, pour ce type de tâche, l'enchaînement des périodes s'aligne prioritairement sur l'enchaînement des actions à décrire.

4. Constructions périodiques et alternance des marqueurs référentiels

Selon le principe d'alternance suggéré dans la section 1, nous nous attendons à avoir plus de marqueurs légers (classe 'CZ') dans les contextes où les périodes fusionnent plusieurs unités de rection et plus de marqueurs lourds ('in' et 'd+') dans les périodes fragmentées. Autrement dit, la nature ± spécifiée des marqueurs serait inversement proportionnelle à la longueur des périodes.

Type de marqueurs / Type de période	Marqueurs lourds (classe IN et D+)	Marqueurs légers (classe CZ)
Périodes alignées	48,84 % (21/43*100)	51,16 % (22/43*100)
Périodes fragmentées	62,5 % (20/32*100)	37,5 % (12/32*100)

Table 1 Corrélations intonosyntaxiques : Nombre de périodes typées avec leurs marqueurs

L'analyse statistique révèle que, pour les périodes qui condensent plusieurs unités de rection, la différence de fréquence d'occurrence des marqueurs lourds et légers n'est pas significative (test de Wilcoxon, $z=-0.152$, $p>.05$). En revanche, elle est significative dans le cas des périodes fragmentées où les marqueurs lourds prédominent (test de Wilcoxon, $z= -3.608$, $p<.05$).

5. Conclusion

Nous avons montré comment il est possible a posteriori de corréler la segmentation en périodes fondée sur des critères exclusivement acoustiques à des contraintes sémantico-pragmatiques sur le marquage de la référence dans le discours, ici une tâche de narration d'histoires en séquence.

L'analyse statistique a révélé que les contraintes sur la distribution des marqueurs référentiels au sein des périodes étaient à géométrie variable en fonction de leur nature, le locuteur privilégiant systématiquement les marqueurs lourds dans les périodes fragmentées. Dans les périodes qui condensent plusieurs unités de rection, en revanche, l'hypothèse selon laquelle les marqueurs légers sont dominants n'est pas

[5] Où les chevrons indiquent les frontières de période.

vérifiée. Sans doute car cette hypothèse est trop simplificatrice : ce ne serait pas tant la nature des marqueurs (léger ou lourd) pris isolément au sein de la période qui serait en jeu mais plutôt le fait que deux marqueurs de même niveau d'accessibilité ne pourraient coexister au sein d'une même période.

Reste à savoir si ces constructions sont également corrélées à la complexité syntaxique des structures en jeu ainsi qu'aux prédicats verbaux manipulés. Enfin, nos données étant semi-spontanées, il serait utile d'étendre nos investigations à d'autres genres oraux pour proposer des généralisations descriptives. Quoi qu'il en soit, nous pouvons supposer que cet équilibre à trouver entre la nature des périodes et le type de marqueur utilisé, qui reposerait sur un principe pragmatique de coopération en lien avec la construction du savoir partagé entre les sujets, est une contrainte invariante quel que soit le genre de discours.

Remerciements

Nous tenons à remercier Mathieu AVANZI pour sa participation dans la transcription et l'annotation des données, ainsi que les personnes ayant participé à cette étude.

Bibliographie

Ariel, M.(1990). *Accessing noun phrases antecedents*, London, Routledge

(2010), Defining Pragmatics.Cambridge university press, (ed).

Avanzi, M.; Lacheret-Dujour, A. & Victorri, B. (2008). ANALOR. A Tool for Semi-Automatic Annotation of French Prosodic Structure. *Proceedings of Speech Prosody'08'*, pp. 119-122.

Bamberg, M. G. (1987). *The acquisition of narratives*, New Babylon, De Gruytey.

Chafe, W., (1974). Language and consciousness. *Language* **50-1**, pp. 111-133.

(1998). « Language and the Flow of Thought ». *The New Psychology of Language*,M. Tomasello (éd.), New Jersey, Lawrence Erlbaum Publishers, 93-111.

Champagne-Lavau, M.; Fossard, M.; Martel, G.;

Chapdelaine, C.; Blouin, G.; Rodriguez, J. & Stip, E. (2009). Do patients with schizophrenia attibute mental states in a referential communication task*?* *Cognitive Neuropsychiatry* **14**(3), pp. 217-239.

Clark, H. H. & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition* **22**, pp. 1-39.

Colle, L. ; Baron-Cohen, S. ; Wheelwright, S.; van der Lely, H. K. (2008), Narrative Discourse in Adults with High-Functioning Austim or Asperger Syndrome. *Journal of austim and developmental disorders* **38**, pp 28-40.

Cornish, F. (2010). Anaphora text-based or discourse-dependent?. *Functions of Language* **17:2**, pp. 207–241.

Gonzalez, S.; Achim, A.; Lavoie, M.-A.; Sandoz, M.; Champagne-Lavau, M. & Fossard, M. (2011), *The Storytelling in sequence test: Assessing Theory of Mind through discourse production.* Third scientific Meeting of the Federation of the European Societies of Neuropsychology, pp. 37-38.

Gundel, J. K.; Hedberg, N. & Zacharski, R. (2012), Underspecification of Cognitive Status in Reference Production : Some Empirical Predictions. *Topics in Cognitive Science* **4**, pp. 249-268.

Lacheret, A. & Victorri, B. (2002). La période intonative comme unité d'analyse pour l'étude du français parlé : modélisation prosodique et enjeux linguistiques*. Verbum* 1-2 **24**, pp. 55-72.

Smith, S. W.; Noda, H. P.; Andrews, S. and Jucker, A. H.(2005). Setting the stage: How speakers prepare listenersfor the introduction of referents in dialogues and monologues. *Journal of Pragmatics* **37**, pp. 1865-1895.

Gender Differences in the Phonetic Realization of Semantic Focus

Carolin Schmid / Sylvia Moosmüller

carolin.schmid@oeaw.ac.at, sylvia.moosmueller@oeaw.ac.at
Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria

Abstract

The aim of the present (pilot-) study is to examine the phonetic focus realization in Standard Austrian German (SAG), with respect to gender specific aspects. Four female and four male speakers were recorded while carrying out diverse dialog tasks, the audio signal was analysed manually. The results demonstrate that the narrower the focus, the stronger is the enhancement of duration, intensity and f0-slope. In addition, an earlier f0-peak is realized on the focus constituent. As concerns gender, differences in the strength of these parameters and in the use of special focus contours were observed. In narrow focus, male speakers tend to use higher values in f0 and intensity than female speakers, who prefer to use steeper or longer f0-falls. In female speech, unlike in male speech, an early high peak often even before the focused word is predominant in narrow focus.

1. Introduction

Semantic focus refers to a linguistic unit which is highlighted within an utterance. Focus marking comprises a semantic-pragmatic structuring function in communication, by highlighting an element containing the informative part of the utterance (see Halliday 1967). In the literature, there is agreement about distinguishing different domains and types of focus. According to the context, the sentence in (1) can contain different focus-conditions, ranging from broad to narrow and from less contrastive to more contrastive. In this study, we will differentiate broad focus (bf), which concerns a larger domain within an utterance, and narrow focus (nf), which is restricted to one word or even only to a part of it. Additionally, within nf, we will distinguish between presentational (npf) and

(1) Katrin went to Italy.
(2) a. What happened?
b. Where did Katrin go?
c. Katrin went to France?

corrective focus (ncf). The context of (2a) introduces the whole sentence (1) as all new information. The context of (2b) elicits a narrow and contrastive object focus by highlighting only one single word. Compared to (2b), the response-alternatives in the context of corrective focus (2c) are smaller and the contrastivity is higher. In unmarked utterances of Indo-European languages, a broad focus is located at the end of an utterance with syntacticly unmarked structure (see Daneš 1974). In marked utterances, however, focus realization is language specific. Büring (2009) describes German as a mixed language, in which focus can be realized either via syntactic or via prosodic means. In this study we will concentrate on the prosodic focus marking realized by the strengthening of the acoustic parameters of an element which is not at the syntactically prominent position of the utterance. Pitch height, intensity, duration, as well as steepness of the rising and falling pitch-contour on the focus-constituent are found to be gradually increasing from broad to narrow and to corrective focus (see Braun & Ladd 2003; Baumann et al. 2006). The timing of the focus-peak is also shown to be influenced by the focus-condition, early peaks indicating a more contrastive focus (see Hanssen et al. 2008). Baumann et al. (2008) note categorical differences between focus-conditions in the use of pitch accents. Some of these authors observed speaker-

specific differences in the weighting and in the strength of the focus-parameters. In studies on intonation, variation has been investigated in consideration of gender as an independent variable. For several varieties of English, it could be shown that High Rising Terminal (HRT)- intonation is influenced by gender (see Warren 2005; Barry 2007). Clopper and Smiljanik (2011) demonstrated that there are gender-preferences for some specific contours within the two varieties of Southern and Midland American English. To our knowledge however, there has been no study so far investigating whether the variation observed in phonetic focus realization is due to social factors or whether it is speaker-specific. Moreover, prosodic focus marking has been investigated primarily for single languages. Some dialect-comparing studies about intonation differences were conducted, implicating nucleus-analyses, which however didn't take into account focus-size and -type (see Gilles 2005; Barker 2005). In the present study, we concentrate on phonetic focus marking in SAG. In particular we are interested whether gender-specific realization of focus will emerge.

2. Method

2.1. Materials

Dialogs were recorded in different tasks:
1. Reading a prepared question-answer catalog to a picture story, so as to elicit exactly identical and therefore highly comparable speech material. As shown in (3), four sentences in respectively three focus-conditions were produced twice per speaker. The focus was always placed in the middle of a sentence to avoid potential influence of sentence -initial and -final pitch movements. The form of a cloze text (speakers had to add the text shown in parenthesis) should mask the real task of focus-production in order to obtain a more natural realization.

2. Playing a dialog-game with animal-cards, in order to evoke ncf and npf. They had to

(3) bf Was ist passiert?
'What happened?'
In einer Geschichte (wollte ein Sohn)
Bücher haben.
'In a story a son wants to have books.'
npf Wer möchte die Bücher haben?
'Who wants to have the books?'
In der Geschichte (möchte der Sohn)
die Bücher haben.
ncf Der Vater möchte die Bücher haben.
'The father wants to have the books.'
In der Geschichte (möchte der Sohn)
die Bücher haben.

guess for each card what animal the partner sees on his upper card, the latter correcting this supposition using ncf (4). Then they had to ask open questions evoking npf (5).

(4) - Sie haben eine Katze vor sich liegen.
'You have a cat in front of you.'
- Ich habe einen Tiger vor mir liegen.
'I have a tiger in front of me.'
(5) - Welches Tier sehen Sie vor sich liegen?
'Which animal do you see in front of you?'
- Ich habe einen Tiger vor mir liegen.
'I have a tiger in front of me.'

2.2. Subjects and procedure

We recorded four female and four male native speakers of SAG (see Moosmüller (1991) for the classification of speakers as standard-speakers). For each recording two speakers of the same sex, who knew each other, had to carry out the diverse dialog tasks. The two speakers were recorded respectively in a separate, closed booth, in order to prevent the additional use of language-external factors. The two booths were connected among each other by microphones and headphones.

2.3. Acoustic measurements

All acoustic measurements have been carried out with STx (see Noll et al. 2001). Utterance-, word- and syllable-segmentation were made manually, as well as the extraction of the relevant phonetic

parameters and the classification of pitch-contours (based on automated pitch tracks calculated using auto-correlation). We measured f0 in terms of mean f0 of the paragraph and the sentence, prominent peaks, as well as valleys before and after the focus-peak. For these valleys, the timing within the sentence was also measured in order to describe the duration of the rising and falling of the pitch around the focus-peak. Additionally we measured the timing of the focus-constituent, more precisely the beginning and the end of the focus-word and of the prominent syllable (so as to find the timing of the focus-peak within the focus-word and -syllable) and the duration of the whole utterance. Also the intensity of the focus-peak, of the surrounding minima and next maxima, as well as its mean value over the utterance was measured. Normalization of duration (indicated as percent of the sentence length), and of f0 values (converted into semitones) was applied. The annotation of pitch contours was based on the GToBI framework (see Grice & Baumann 2002). "<" shows that the focus-peak is realized before the focused syllable.

3. Results

3.1. F0

We measured the mean height of the focus-peak, relative to the mean sentence frequency. Figure 1 shows the peak-values,

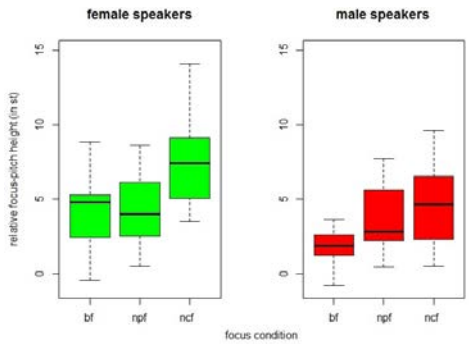


Figure 1: Height of the focus-f0-peak (relative to mean f0)

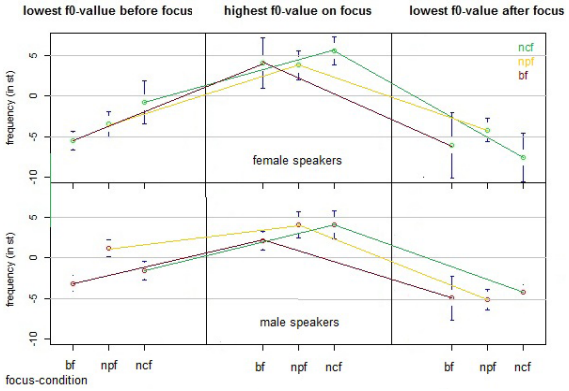


Figure 2: F0-rise and -fall around the focus-peak

separate for gender and focus-condition. The narrower the focus is, the more the relative height of the focus-peak is increasing. Nevertheless we observe a different grouping, insofar as male speakers don't realize a significant peak-difference between npf and ncf, whereas female speakers rather group bf and npf together and enhance the production of the ncf-peak. The mean sentence-f0 is decreasing with narrower focus in both female and male speech. Figure 2 suggests that the f0-fall from the focus-peak to the following valley in both male and female speech is more extreme than the rise from the precedent valley to the focus-peak. In female speech, these differences are stronger than in male speech. Finally, we also analyzed the timing of the f0-peak within the focus-word. Preliminary analyses of the pitch-transcriptions (so far only for nf) reveal that women make more use of early peak-contours than male speakers (see Figure 3). Often, the early peak lays even before the focused word and precedes a longer and steeper pitch-fall.

3.2. Duration

Concerning duration patterns, we couldn't find any difference in the realization of the word-duration both for focus condition and for gender. However, duration of the focus-syllable seems to be correlated with focus condition, and also shows gender-specific characteristics. Figure 4 shows that the duration on the focus syllable is longer than the mean syllable duration, and the latter is

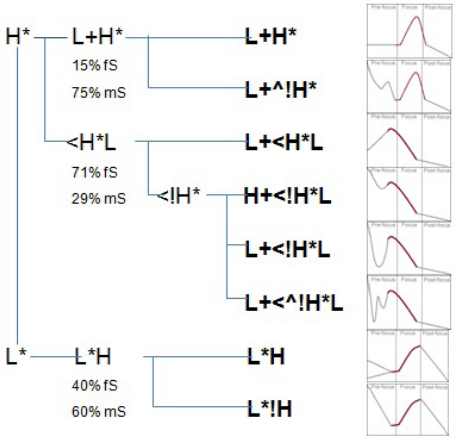


Figure 3: Occurence of focus-contours in nf (fS=female speakers, mS=male speakers)

decreasing with narrowing focus. While in female speech, mean syllable duration and focus syllable duration decrease in equal measure over all focus conditions, the focus syllable duration in male speech decreases to a lesser extent than the mean syllable duration, thereby creating an increase of the focus-syllable duration relative to the mean syllable duration in narrowing focus. Duration of the pitch movement on the focus peak (rising and falling) is always longer in the falling movement than in the rising movement. In female speech, the duration of the falling contour is increasing the narrower the focus is. As demonstrated in Figure 5, duration of the pitch-movement on the focus peak is clearly increasing with narrowing focus in female speech. In male speech, on the other hand, duration is longest in bf, whereas in both nf-conditions, an even shorter pitch movement on the focus peak can be observed.

3.3. Intensity

The relative intensity on the focus peak is higher in each focus condition than the mean intensity of the sentence and is also increasing with narrowing focus (as shown in Figure 6). Male speakers tend to contrast bf vs. npf more than npf vs. ncf, whereas in female speech, there is a tendency to more contrast between ncf on the one hand and bf and npf on the other. In male speech, the

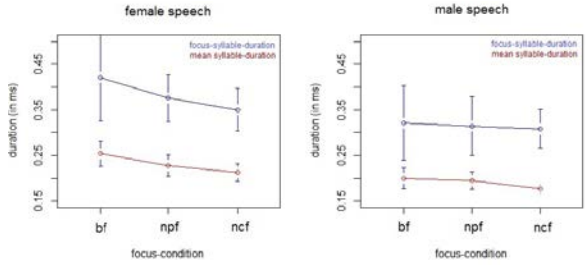


Figure 4: Mean syllable duration vs. focus-syllable duration.

contrast between focus-intensity and mean intensity of the sentence is higher than in female speech. Our results also indicate that, relative to the precedent and following intensity-peaks, intensity shift on the focus word is extremer in nf-conditions, the intensity fall after the focus word being stronger than the preceding rise. Female speakers distinguish more between all focus conditions whereas male speakers group npf and ncf together, but realize stronger intensity-shifts.

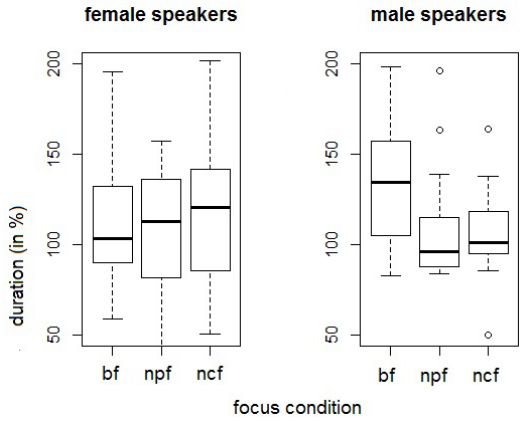


Figure 5: Duration of the pitch movement to and from the focuspeak (in percent of the focus-word duration)

4. Discussion and Conclusion

In this pilot study we found evidence for an acoustic differentiation between the focus conditions bf, npf and ncf. However, a categorical differentiation did not evolve: Both nf conditions are often merged, or npf is realized as bf. Nevertheless, we observe gender-specific grouping of the focus-conditions. In female speech, bf and npf are more often grouped together by relatively enhancing ncf. This becomes obvious both in f0-peak production and in the realization

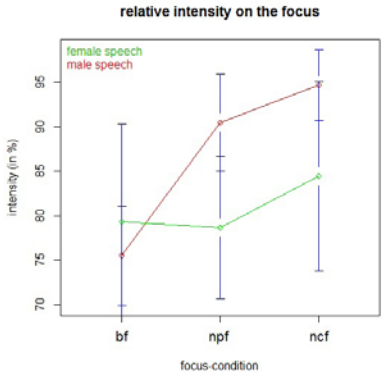


Figure 6: Highest intensity on the focus-word (in percent of the range between mean and highest sentence intensity)

of the highest focus-intensity. In male speech on the other hand, we found a grouping of npf and ncf, of the realization of the f0-peak, of the duration of the pitch movement, and also of the highest focus-intensity. The investigated parameters f0, duration, and intensity are all found to contribute to phonetic focus production, and we also observed gender-specific preferences in the use of the investigated focus parameters. Female speakers use more extreme relative f0-differences on the focus to signal the focus-word. We also found differences in the distribution of focus-pitch contours. Women prefer to produce earlier peaks preceding a stronger fall on the focus-word, while men realize f0 peaks more often within the focus-word. The duration of the pitch movement around the focus is increasing with narrowing focus in female speech, while in male speech it is the other way round. Analyses of duration showed that syllable duration is used more in male speech than in female speech to differentiate the focus conditions. Male speakers also make more use of intensity to encode focus-meaning than female speakers. The results suggest that at least some of the variability found in focus realization can be ascribed to gender-specific differences in focus production. However, more research on this topic is needed to obtain significant and representative results. Additional speaker-recordings are currently underway to test the results obtained so far.

References

Barker, G. (2005). *Intonation patterns in Tyrolean German: An autosegmental-metrical analysis*. Peter Lang, New York.

Barry, A. S. (2007). *The form, function, and distribution of high rising intonation in Southern Californian and Southern British English*. University of Sheffield.

Baumann, S., M. Grice & S. Steindamm (2006). Prosodic Marking of Focus Domains - Categorical or Gradient?. *Proceedings of the 3th International Conference on Speech Prosody*, pp. 301-304.

Braun, B. & D.R. Ladd (2003). Prosodic Correlates of Contrastive and Non-Contrastive Themes in German. *Proceedings of 8th European Conference on Speech Communication and Technology*, pp. 789-792.

Büring, D. (2009). Towards a Typology of Focus Realization. Zimmermann, M. & C. Féry (eds.), *Information Structure.* Oxford University Press, Oxford, pp. 157-205.

Clopper, C.G & R. Smiljanik (2011). Effects of gender and regional dialect on prosodic patterns in American English. *Journal of Phonetics* 39:2, pp. 237-245.

Daneš, F. (1974). Functional sentence perspective and the organization of the text. Daneš, F. (ed.), *Papers on Functional Sentence Perspective.* Academia, Prague, pp. 106-128.

Gilles, P. (2005). Regionale Prosodie im Deutschen: Variabilität in der Intonation von Abschluss und Weiterverweisung. De Gruyter, Berlin/New York.

Grice, M. & S. Baumann (2002). Deutsche Intonation und GToBI. *Linguistische Berichte* 191, pp. 267-298.

Halliday, M.A.K. (1967). Notes on transivity and theme in English, Part 2. *Journal of Linguistics* 3, pp. 199-244.

Hanssen, J., J. Peters & C. Gussenhoven (2008). Prosodic Effects of Focus in Dutch Declaratives. *Proceedings of the 4th International Conference on Speech Prosody*, pp. 609-612.

Moosmüller, S. (1991). *Hochsprache und Dialekt in Österreich. Soziophonologische Untersuchung zu ihrer Abgrenzung in Wien, Graz, Salzburg und Innsbruck*. Sprachwissenschaftliche Reihe 1, Böhlau ,Wien/Köln/Weimar.

Noll, A., Deutsch, W. A., Balazs, P., White, J. (2001): *Intelligent Sound Processing S_TOOLS - STX User's Guide*.

Warren, P. (2005). Patterns of late rising in New Zealand English: Intonational variation or intonational change?. *Language Variation and Change* 17:2, pp. 209-230.

Large-scale analysis of call centre conversations: call structure as prosody

Rein Ove Sikveland & David Zeitlyn

rein.sikveland@anthro.ox.ac.uk, david.zeitlyn@anthro.ox.ac.uk
Institute of Social and Cultural Anthropology, University of Oxford

Abstract

Based on findings in conversation analysis (CA) we develop methods for studying conversational structure, prosody and 'success', for potential use within management of call centres. Our research combines CA knowledge with computational tools and corpora methods. In this paper we present the objectives of our research, while highlighting the challenges involved in combining different research methods, and automating the analysis of conversations.

1. Introduction

1.1. Objectives

In our research we seek to automate the identification of unusual or problematic interactions in a large corpus of call centre conversations. In an applied perspective this may lead to improvements in the everyday management of call centres.

Although we have access to call centre quality control identification of problematic calls (i.e. those which (human) managers judge to be problematic relative to the normal flow of calls in the course of doing their work, not we note motivated by criteria of scientific research and consistency) we are trying in the first instance to identify unusual calls which are outliers in terms of their conversational structure. Our working hypothesis is that many such outliers will also be judged problematic.

We aim to develop ways of extracting important elements of conversational structure, such as who talks when and in what relation to next or previous turns in a conversation. In a speech technological perspective this shifts the focus from the semantic content of words and phrases, to how interactions are maintained and organised.

We focus on prosodic markers of discourse function as well as structural elements of conversations (e.g. turn-taking), combining corpus linguistic and quantitative techniques with a conversation analytic (CA) approach (see Sikveland & Zeitlyn, submitted; Kurtic et al. 2013).

Scaling up the analysis of conversations and discourse is an important challenge in our research, and is a form of the 'big data' challenge which is currently being addressed in a wide range of research. In this paper we illustrate why this is an interesting methodological challenge, and ways of addressing it. We also aim to demonstrate the value of combining very different methodological approaches to offer a new angle towards the study of discourse and prosody.

1.2. Tools and terminology

We seek to develop means to automatically identify unusual or interesting events in a (large) set of conversations. In other words, we target particular characteristics of talk-in-interaction that the call centre management can address further. Our building blocks for answering these questions are: 1) large amounts of call centre material, 2) computational tools and algorithms for identifying speech, speaker and particular elements of speech production, 3) input as to what call centres are looking for in a call, 4) knowledge/background in conversation analysis and prosodic analyses.

The analyses we present here focus on conversational structure, in particular durational sizes of turns and pauses in terms of their distribution, and in terms of the relationships between consecutive turns. Rather than studying individual utterances, or utterance forms, we study a wider discourse structure – the overall structure of a conversation. 'Conversational prosody' might be a fitting term for our focus: our definition and use of 'prosody' goes beyond features like pitch, intensity and speech rate/rhythm, and may include any aspect of communication which is not traditionally described as part of the linguistic system (cf. Ogden 2012, on 'prosodies'; see also Couper-Kuhlen & Ford (2004) for an introduction to studies of turn-taking and interactional linguistics/prosody).

2. Background

2.1. Action and interactional sequence

Our research is inspired by CA work, an area of language research which has distinctly and explicitly addressed the distinction between 'information' and 'action' (Schegloff 1995). There is often no direct (simple) relationship between form and function: CA argues for involving extended sequential properties of conversations to better formulate such a relationship. For example, the form of request may depend on where in a conversation it is used and what kind of response it seeks (Curl & Drew 2008). Furthermore, prosodic display of agreement is also tightly linked to sequential properties (Ogden 2006). By examining sequence we can demonstrate how contextual factors shape language, and conversely, how language shapes context; and how linguistic and other communicative resources are used according to this process.

2.2. Presence and absence

The interactional process is just as much about the absence of events as it is about the presence of particular content. The example below illustrates this both in terms of pausing and prosody. Importantly, the absence of a particular event, e.g. a display of agreement, is not an absence of relevant information regarding the action in process. The notion of presence/absence is a cornerstone in CA research (see e.g. Sacks 1992, Vol.1: 294). Inspired by such a framework, we aim to treat units of talk as part of a larger activity. Consequently, the duration and distribution of pauses is related to the organisation of talk and action. Absences (silences) are consequential and might indicate trouble. This is the starting point for our research.

Participants' own orientation to absences in illustrated in Example 1 below: Chris (C) has just been telling Wendy (W) about recent achievements at the school where he works. In 01 W forms an assessment in response to this, which continues until the end of 08.

(1)

01	W:		it's uh (1.4)
			it's quite an event and you're lucky that you're
			there at that time you know hh=
02	C:		=right .h[h right]
03	W:		[you get the] mor:e
			press: and stuff for the school itself
04	C:		yeah
05			(.)
06	W:		which is good nhh
07		->	(.)
08	W:	->	.tk .hhh so [that's a] good thing
09	C:		[.thh]
10			(0.5)
11	C:	->	yeah mhhhh
12	W:		(and that) well the one
			important thing I'd better ask you - - -

Note that although C responds verbally in lines 02 and 04, these responses do not explicitly agree with or upgrade W's assessment. C is merely aligning with the talk in progress, and based on observable events we may argue that there is an absence of talk from C. Firstly, W orients to this absence by incrementing on her own talk: *which is good* in 06 does not add anything to

her previous assessment, except bringing it forward, potentially to a topic closing. The following gap of non-speech (07) is clearly consequential to the progressivity of the current talk: W repeats and upgrades her previous increment in 08. This repeat makes relevant a more explicitly formed agreement from C: although lexically C's *yeah* has the same form in 04, prosodically the second *yeah* (11) is louder, has higher average pitch and larger pitch movement than the first one.

Knowing that this is an assessment sequence we might note the duration of gaps and the form and placement of responses as 'not standard': The response in 04 is short and prosodically 'weak', the second one response (07) is missing, and the third response (11) is a delayed upgrade of the previous response.

Can we use similar features to detect 'non-standard' sequences elsewhere? In a long term our success in formulating a model which may automatically detect the presence (and form/function) of a response, as well as their absence, requires a means of detecting the type of structure it occurs in.

2.3. A (bottom-up) corpus linguistic approach

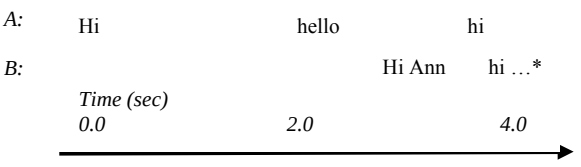
One way of approaching our research interests would be to start with a top-down model of action and sequence, based on manually labelled activity types and sequential relationships. However this is both expensive and challenging: the labelling would need to be based on detailed analyses of both content and interactional processes, which might not be practical (or suitable) for large amounts of data (see however Shriberg et al. 1998 for example of large-scale corpora with dialogue acts labelled).

We chose a different approach: since we do not have content data, we started with the conceptually straightforward task of automating the detection of speech chunks (i.e. who talks when). Thus we deal with a rudimentary structure, not directly connected with linguistic content.

Rather than being a problem, the absence of content can be turned to our advantage – particularly as we are working with large quantities of data. For example, we can ask whether there are features that go along with particular sections of a conversation, and whether, based on these features, we can separate calls into sections and thereby compare the overall structure across calls.

The opening of the example phone call below contains a rather long gap between the greeting (sp. A) and response to the greeting response (sp. B):

(2)



This example is taken from an actual call centre telephone opening, where the participants themselves show an absence of relevant response following A's first *hi*: following the final *hi* B adds *I took a swallow there (laugh) while answering*. This is a non-standard opening of a call. By (i) operationalising a rudimentary call structure, and (ii) comparing single calls with the structure of a large aggregate of calls we aim to identify such unusual events in calls. The techniques we use include the training of Hidden Markov Modelling (HMM), in order to automatically detect structures based on a sequence of events.

3. Discussion

3.1. Bringing together different traditions

From the outset it might be difficult or controversial to bring together two different research traditions. In our case we combine one method which is sometimes a critique of the other: for example, coding and

quantification is problematised within a CA framework (Schegloff 1993).

Karen Tracy and Julien Mirivel (2009: 159) are clear about the differences in approach between the qualitative analysis of a limited number of cases, and quantitative analysis relying on the coding of interactions. Coding research might on the one hand be criticised for the lack of attention to the details involved in participants' own local management of turn-taking and interaction. However, where the coding is arrived at inductively on the basis of CA-inspired qualitative work, they see a relationship of potential friendship between the two very different methods. We too are working in a spirit of amity and friendship inspired by CA yet the scale of data we want to consider means we need to use systematic and statistical approaches: however, our methods do not rely on a coding scheme in the usual sense of the term.

The potential benefits of automating analysis are great, but the challenges are many, and we need to carefully consider what sorts of analyses and applications are possible to achieve. At the centre of our challenge stands the operationalisation of not only a collection of single features, but also of the structures and structural relationships in which those features are embedded.

A combination of methods does not necessarily mean an improvement in our qualitative understanding of a phenomenon, e.g. of turn-taking (Tracy 1993). However it can aid us in approaching conversational data in new ways and discovering new knowledge regarding their structures.

3.2. How are 'interesting data' discovered?

We approach here the question of what constitutes 'interesting', or 'unusual'. Since we need to deal with more data than a single group of humans can listen to we are faced with challenges which conventional Conversation Analysis has politely avoided discussing. When we started to operationalise forms of approaches that rested on CA principles we realised that there were unexamined assumptions in CA which could not easily be made explicit. These concern the identification of the cases to be analysed, in effect resting on what constitutes an 'interesting' or 'problematic' case.

In CA it is maintained that there is no particular or general motivation for selecting a particular case (Sacks 1992: 293; Sacks adopts a counter-strategy to picking 'interesting' data by focussing on *a priori* 'un-interesting' data, e.g. openings of phone calls). From a CA point of view something interesting can be found in any piece of data; however this stands in contradiction to another starting point in CA, namely that analysis should not be motivated by previous knowledge and expectations. The notion of '(un)interesting' necessarily builds on a rich understanding of the language and the situation. Thus, although the emerging analysis can have strongly objective qualities, there is clearly an element of a nativeness, grounded in unspoken native speaker intuitions, that governs the work (Fitch 2005: 472).

Though this need not be a problem in itself, or for individual research projects, it leaves a number of questions unanswered. For example, has proper attention been given to exceptions? What are they and how frequent are they? The notion of commonly used practices must imply that there are unusual ones: what are they and in what circumstances do they occur - in other words, how can they be detected?

In his introductory textbook Paul ten Have (2007: 120-126, 140-142) explains how important for CA is the use of 'unmotivated looking' but he does not explain how interesting cases are identified (similarly vague are Jordan & Henderson 1995; Hutchby & Wooffitt 2008: 26.89; see also Schegloff 1999: 577-578).

Introductory works on CA typically talk

of a three stage process: recording, transcribing and data sessions. There is a lot of practical advice available about the first two, but surprisingly little information about how to run a 'data session' let alone how to choose the instances which will be exhaustively analysed. This has been a problem for us since we have been looking at ways to deal with large streams of data. If we cannot sift out routine, unproblematic instances of conversation to latch onto the interesting or problematic then we will be irremediably stuck in the situation of data deluge with which we began; and it would seem that CA cannot help us overcome it.

3.3. Conclusion

We have been considering what kinds of interactional analysis are possible without recourse to content. Is this at all possible, and how can we effectively include linguistic and extra-linguistic content in automated interactional analyses in the future? What are the particular methodological challenges involved?

Inspired by Tracy and Mirivel's (2009) promise of friendliness we have proceeded on the basis of such amity and are undertaking a large scale and statistical yet CA inspired analysis of our datasets.

Our aim is to automate the study of multiple calls using a set of features: we use a quantitative approach to form an analysis of commonalities across conversations; and thereby also what may constitute an unusual event. Although our current work focuses mainly on durational features, we see it as a starting point for work including other prosodic, as well as lexical and syntactic properties in the future.

Acknowledgments

We wish to thank the Knowledge Transfer Partnership (KTP), and the UK Technology Strategy Board (TSB), which funded this project.

References

Antaki, C., M. Biazzi, A. Nissen, J. Wagner (2008). Managing moral accountability in scholarly talk: the case of a Conversation Analysis data session. *Text and Talk* 28, pp. 1-30.

Couper-Kulhen, E. & C.E. Ford (2004). *Sound Patterns in Interaction*. Amsterdam: John Benjamins

Curl, T.S. & P. Drew (2008). Contingency and Action: A Comparison of Two Forms of Requesting. *Research on Language and Social Interaction* 41:2, pp. 129-153.

Fitch, K.L. (2005). Conclusion: Behind the Scenes of Language and Scholarly Interaction. K. L. Fitch & R. E. Sanders (eds.), *Handbook of language and social interaction*. Mahwah, NJ: Erlbaum, pp. 461-482.

Kurtic, E., G.J. Brown, B. Wells (2013). Resources for turn competition in overlapping talk. *Speech Communication* 55, pp. 1-23.

Ogden, R. (2006). Phonetics and social action in agreements and disagreements. *Journal of Pragmatics* 38, pp. 1752-1775.

Ogden, R. (2011). Conversational prosody. O. Niebuhr (ed.), *Understanding Prosody.* Göttingen, Germany: De Gruyter, pp. 201-218.

Sacks, H. (1992). *Lectures on Conversation, Volumes I & II*. UK: Blackwell Publishing.

Schegloff, E.A. (1995). Discourse as an Interactional Achievement III: The Omnirelevance of Action. *Research on Language and Social Interaction* 28:3, pp. 185-211.

Schegloff, E. A. (1999). Naivete vs sophistication or discipline vs self-indulgence: a rejoinder to Billig. *Discourse and Society* 10:4, pp. 577-582.

Shriberg, E., R. Bates, A. Stolcke, P. Taylor, D. Jurafsky, K. Ries & D. Van Ess-Dykema (1998). Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech* 41:3-4, pp. 443–492.

Sikveland, R., & D. Zeitlyn. Combining conversation analysis and corpus linguistics: prosodic cues alone do not identify dialogue acts. *International Journal of Corpus Linguistics* (submitted).

Tracy, K. (2005). Reconstructing communicative practices: Action-implicative discourse analysis. K. L. Fitch & R. E. Sanders (eds.), *Handbook of language and social interaction*. Mahwah, NJ: Erlbaum, pp. 301-319.

Tracy, K. & J.C. Mirivel (2009). *Discourse Analysis*. L. R. Frey, K. N. Cissna (eds.), *Routledge Handbook of Applied Communication Research*. NY: Routledge, pp.153-177.

Marking boundaries: intonation units and prosodic sentences

Candide Simard

cs75@soas.ac.uk
School of Oriental and African Studies, London, UK

Abstract

This paper presents a study of the prosodic correlates at the boundaries of intonation units and prosodic sentences in Jaminjung, a severely endangered language of Northern Australia, including pitch resets, final lengthening, pauses, and phonation events such as breathiness and creakiness. This analysis demonstrates that units of speech larger than IUs must be examined to account for the phenomena observed in spontaneous speech. It contributes to the on-going debate on the nature and status of discourse units and the best methodologies for their identification (Degand and Simon 2009); it also contributes to the developing interest in examining discourse in spontaneous speech and its theoretical implications (Wichman 2006) from the perspective of an, as yet, unwritten language.

Keywords: prosody, discourse, intonation units, prosodic sentences, boundaries, Australian languages, Jaminjung.

1. Introduction

Speech is not uttered as a monotonal unbroken string, it is organised into units of various sizes which group together to make larger units expressed through prosody (mostly) (Swerts & Geluykens, 1994).

Phonological models of intonation usually restrict their analysis to that of the intonation unit (IU hereafter), cued phonetically by a coherent pitch contour and bounded by pauses, roughly corresponding to a clause (Chafe 1994, Ladd 2008 inter alia). Patterns spanning segments larger than the IU are viewed as discourse related. In prosodic analysis, larger units have been described as paratones (Crystal 2003); as paragraphs by Lehiste (1975) and Tseng (2010); as spoken sentences by Wichman (2000: 128), and Genetti (2007) refers to prosodic sentences in her study of Dolakha Newar. While we do not claim that these terms are used synonymously, it is clear that researchers agree on the usefulness of a unit larger than the IU to account for the phenomena observed in speech. We will adopt the term prosodic sentence (PS), following Genetti, because it implies a cohesion between a number of units. However, PSs are not always co-extent with syntactic sentences. Consider a list of NPs that would not constitute a syntactic sentence but could easily make a PS (see Genetti, 2007 for more relevant examples). Also, spontaneous speech is full of disfluencies, such as repairs, repetitions, filled pauses, etc. We will examine the differences in the prosodic encodings of IUs and PSs in a language spoken around Timber Creek in the Northern Territory of Australia, the remaining few dozen speakers of which are all elderly. The analysis of the prosodic constituents in Jaminjung, IUs and PSs, is part of its on-going description, thus providing direct input for our understanding of its morpho-syntax, semantics and pragmatics. It provides insights about the expression of the dependency relations between its clauses, semantic cohesion (topicality, parenthesis, afterthougths, etc) and interactional management.

Jaminjung is part of a small western branch of the geographically discontinuous Mirndi family, a member of the diverse non-Pama-Nyungan group of Australian languages (Harvey 2008). As in other Australian languages, Jaminjung is said to have ‘free word order’ in that the ordering of words does not indicate the grammatical roles of arguments, but it is rather conditioned by information structure at the discourse pragmatic level (Schultze-Berndt 2000), thus our interest in investigating it more closely. This leads to lexical arguments being

optionally omitted – although argument roles are indicated by bound pronominals which attach to the verbs as prefixes and by case markers suffixed to constituents of noun phrases. Like many other local languages (McGregor 2004), Jaminjung has two distinct categories of verbs: inflecting verbs, which form a closed class of around thirty members, and a non-inflecting open category, referred to as 'coverbs' (Schultze-Berndt 2000).

In the model of prosodic analysis adopted here, the Parallel Encoding and Target Approximation (PENTA) model (Xu 2005), different communicative functions are simultaneously encoded in a single prosodic contour. The prosodic parameters which encode specific functions can be investigated with quantitative and statistical tools. For example, syllable duration has been associated with boundaries of various size; average energy (intensity) with stress; F0 (fundamental frequency of the syllable) with topicality, focus, and contrast; and the slope of the F0 contour (visualised as intonation rising, falling or flat) with sentence types.

The prosodic correlates of the syllables at the edge of IUs and PSs are compared, to show that the two units of grouping can be differentiated on prosodic grounds. At the left edge, pitch resets are measured; at the right, duration, expressed in relative terms, is examined to assess the extent of final syllable lengthening; final pitch lowering is also considered. Lastly, in Jaminjung, speech is sometimes still perceptible after the last calculated point of the pitch track. The aperiodicity in the signal is associated with either breathy or creaky phonation which are hypothesised to be potential boundary cues.

2. Methods

2.1. Dataset

The data is extracted from a corpus of spontaneous speech resulting from fieldwork conducted between 1993[1] and 2009. The dataset consists of four narratives: one mythological story, two personal anecdotes and a picture-prompted retelling of the Frog Story (Mayer 1969) from four different speakers (all female), totalling 213 PSs and 448 IUs. Recordings usually involved more than one speaker. Field-based audio recordings are rarely of optimal quality for acoustic analysis; hence our datasets are limited in number. Moreover, we had to disregard measurements of intensity as it was impossible to control the distance between speaker and microphone during recording sessions. Nonetheless, we contend that the analysis is still worthwhile and feasible: patterns must be identifiable, otherwise speakers would not use them in their interactions.

As the stated aim of this study is to discover the prosodic cues associated with each unit, in order to avoid circularity, the identification of the posited IUs and PSs is based primarily on semantic and syntactic criteria. However, it is acknowledged that the presence of pauses is one of the major criteria for establishing IUs during the first transcription of the data, prior to any analysis. Semantically, a PS is a sequence of related IUs. Syntactically, it may correspond to a succession of independent clauses, a main clause and a subordinate clause, to a main clause and a dislocated element, either an afterthought or a dislocated topic. Occasionally, it also corresponds to direct speech, as an argument of a reporting verb, or to an interjection (forming its own IU) and a main clause.

As a reference, example (1) shows the prosodic contour in a PS from a Frog Story consisting of a single IU made of a topic, the NP *thanthubiyang ngayin* 'those animals' and a comment, *bunburr burrangga* 'all take

[1] Much of the data in the corpus was collected by Dr Eva Schultze-Berndt, who made it available to the author.

off'; the focus is on the first syllable of the comment, in the coverb *bunburr*.

(1)
thanthu=biyang ngayin bunburr burr-angga
DEM=now animal many.take.off 3pl-o.PRS
'Those animals all take off.' [IP:ES97_03_01]

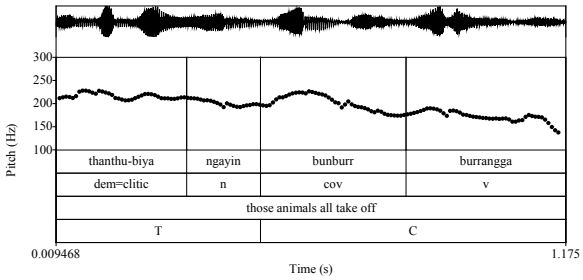


Figure 1 A prosodic sentence made of a declarative IU, with the default falling contour. Tier 1 shows the prosodic words, tier 2 the parts of speech, tier 3 is a free translation and tier 4 a Topic-Comment segmentation.

Example (2) shows a PS made of two verbal IUs, describing a succession of events in a narrative in which the speaker recalls being bitten by a centipede.

(2)
<that> jalarriny <bin> wirriny ngarrgu
that centipede been turn 1SG.OBL
mam gan-ba=biyang
hold.tight 3sg:1sg-bite.PST=now
'That centipede turned on me, and bit me with a tight grip (i.e. it didn't let go).' [IP:ES97_03_02]

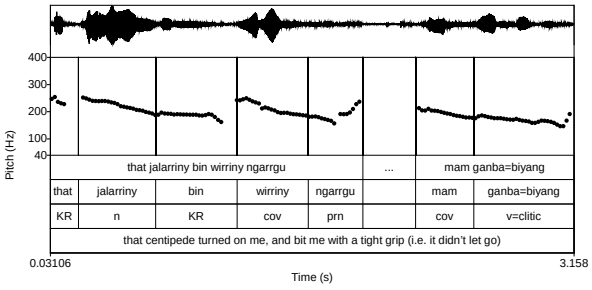


Figure 2. A prosodic sentence containing two IUs, describing a succession of events in a narrative.

Example (3) shows a PS from the same narrative made up of two IUs, the second consisting of an NP *<buj>-mawu buyud* 'the bush-kind of sand' serving as an afterthought.

(3)
buyud=biyang <jabul >-ni burr-angu=rrgu=rndi
sand=now shovel-INSTR 3pl:3sg-get/handle-PST=1sg.OBL=FOCUS

<buj>-mawu buyud
bush-ORIG sand
'They got sand for me with a shovel, the bush kind of sand.' [IP:ES97_03_02]

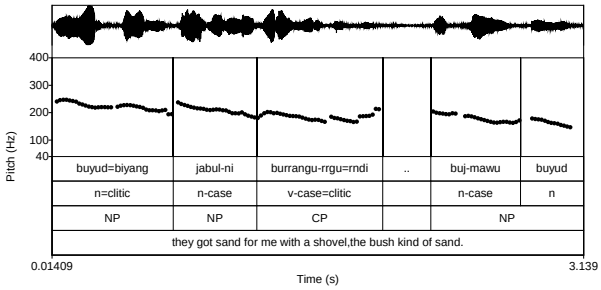


Figure3 The second IU in this prosodic sentence is formed of an NP serving as an afterthought.

2.2. Measurements

The data is analysed and annotated using Praat software (Boersma & Weeninck 2013). All IUs in the dataset are labelled according to their syntactic subtypes and segmented into syllables. The number of words in each IU and their positions are indicated, also the number of syllables and their position in a word. The location of breathiness and creakiness is also noted. Finally, the following measurements are made for each syllable (using a script developed by Xu[2])

- Mean F0 — Average of all F0 values in a syllable (10 measurements), in Hz.
- Excursion size — Difference between the max and min F0 expressed in semitones for each syllable.
- Final velocity — Velocity is a measure of the instantaneous rates of F0 change expressed in semitones per second, taken at a point earlier than the interval offset (here 30ms). It is an indicator of the slope of the underlying target of the interval.

[1] http://www.phon.ucl.ac.uk/home/yi/tools.html.

- Duration — Time interval between the onset and offset of the syllable, in ms.

At the left edge, mean pitch measures are used to calculate the degree of pitch reset, calculated by subtracting the mean F0 of the last syllable of a unit from that of the following syllable, formulated as:

F0 Reset = MeanF0 LastSyll – MeanF0 Next Syll .

At the right edge, measurements of absolute duration are used to calculate final lengthening. The relative duration of the last syllables is established as a ratio of the length of the final syllables compared to that of the preceding syllables in the same unit. For speech sounds (segments), usually ranging in duration from 30ms to 300ms, studies have shown that differences must be 10 to 40ms in length before they are judged by listeners to have a 'just-noticeable difference' in duration (Lehiste 1970:13). This is taken into account in the interpretation of the results below. In order to establish a threshold, we follow Amir et al. (2004) in considering a syllable to be lengthened if it is 10% longer than the preceding syllables (ratio > 1.1). The occurrences and lengths of pauses are counted. Final lowering is also a relative value, comparing the mean F0 of the final syllable to that of the syllables at the beginning of the unit. In our dataset, the presence or absence of breathy and creaky phonation is identified firstly through perception, then using the waveform, pitch tracker, and spectrogram.

The quantitative analysis consists of a comparison of the values for each correlate for the boundaries of IUs and PSs. A statistical analysis is conducted to validate the results. Preliminary tests assess the impact of speaker difference and IU length on the analyses.

3. Results

3.1. Preliminary tests

To ensure uniform treatment of all data, which consists of IUs with differing lengths by different speakers, tests are conducted to assess whether the factors 'speaker' and 'IU length' affect our analyses. Multifactor ANOVA tests show that the interaction between 'subtype' and 'IU length' is not significant for any of the correlates; neither is the factor 'speaker'.

3.2. Left boundary

The calculation of the pitch resets in the first syllables of IUs and PSs shows a significant difference between the two, with averages of 11.21Hz for IUs and 22.47Hz in PSs as shown in the table below.

Unit	Mean pitch reset (Hz)	Std. Dev	N
IU	11.21	37.21	224
ProsS	22.47	52.32	201

3.3. Right boundary

The relative duration of final syllables in IUs and PSs is expressed as ratios with respective values of 1.23 and 1.11, an unexpected result. Continuing the analysis with the occurrence and length of pauses, we find IUs are mostly, but not necessarily, bounded by pauses, occurring in 89% of all tokens. PSs are always bounded by pauses, but as mentioned earlier, this is definitional, and potentially circular, in this study. The duration of pauses after IUs is 689.98ms while those after PSs are much longer, averaging 1653.40ms, a difference that is statistically significant; the standard deviation value also highlights a wide variation in the length of pauses for each unit (F (2, 399) 61.389, p=.000).

The absolute F0 average of final syllables in IUs is 164.22 Hz and 147.69Hz in PSs. It is more useful, however, to view final lowering as a relative value, respectively comparing the first and last syllables in both IUs and PS. The graphs in Figure 4 shows that IUs (left pane) undergo only a slight lowering, which may be the result of averaging the F0 of final syllables in IUs that receive non-terminal continuation rises with those that do not. The

right pane shows the lowering found in PSs, which is much more marked.

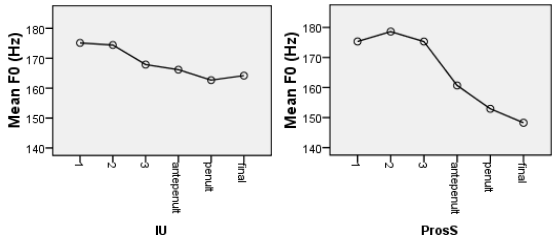


Figure 4 The lowering of F0 shown for each type of unit. The F0 of the first and last 3 syllables of each unit are considered.

Finally, as to phonation types, breathiness occurs predominantly in IUs (71% of occurrences), while creakiness is associated with the boundary of PSs (89%). The correlation between phonation type and type of unit is verified with a Chi-square test which shows a strong correlation (p= .000).

4. Discussion and Conclusion

This paper examined the prosodic correlates at the boundaries of IUs and PSs in Jaminjung. It demonstrated that speakers systematically use prosody as a demarcative device in discourse, segmenting larger chunks of information with longer pauses, low boundary tones and higher pitch resets.
The units are indeed distinguished at their left boundaries by pitch resets which increase from unit to unit, a finding similar to that of Schuetze-Coburn et al. (1991) for American English. This melodic discontinuity between information units is viewed as an important cue for discourse segmentation at least in read or rehearsed speech (Wichman 2000). The results presented here suggest the same appears to be true for spontaneous speech in Jaminjung.
At the right boundary, both units are lengthened. The values of IUs' duration ratios are slightly greater than those of PSs, results which may be explained by the presence in the dataset of IUs with an overall level contour, a contour specific to Jaminjung[3], characterised by markedly long final syllables, possibly associated with the expression of durativity. The pauses following IUs are usually shorter than those following PSs. Final lowering occurs in both IUs and PSs but is much more salient in the larger unit. IUs may end in breathy phonation and prosodic sentences in creaky phonation, a finding that accrues the mounting body of evidence that considers phonation as a likely discourse structuring device (Gordon and Ladefoged 2001). As all the measures in this study display some gradience, it could be argued that non-finality is not marked categorically in prosody (at least not in a binary final/non-final manner).
In conclusion, this analysis of the correlates associated with the boundaries of larger prosodic units in Jaminjung is important in contributing a rare analysis in an Australian language, and thus helps in validating the existing findings from European or better-known languages for which discourse characteristics in general and prosodic features in particular have been studied and which form the basis for universal claims. They also point to positing the right boundaries of a larger unit in discourse, the prosodic sentence, as the locus where syntactic completion, semantic cohesion and pragmatic /interactional management actually coincide in signalling finality.

5. Acknowledgments

We wish to acknowledge the patience and knowledge of the many Jaminjung and Ngaliwurru speakers who have worked with us over the years. We are also grateful for the funding received from the DoBeS programme of the Volkswagen Foundation for the documentation of the linguistic and cultural knowledge of Jaminjung and other languages of the Victoria River district.

[3] And possibly other Australian languages.

References

Amir, N. & V. Silver-Varod & S. Izreel (2004). Characteristics of Intonation Unit Boundaries in Spontaneous Spoken Hebrew – Perception and Acoustic Correlates. In B. Bel & I. Marlien (eds.), *Proceedings Speech Prosody 2004*, Nara, Japan, March 23-26. pp. 677-680.

Boersma, P. & D. Weenink, David (2013). *Praat: doing phonetics by computer [Computer program]*. Version 5.3.42, retrieved 2 March 2013 http://www.praat.org/.

Chafe, Wallace (1994). *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago: The University of Chicago Press.

Crystal, D. (2003). *A dictionary of linguistics and phonetics*. Oxford: Blackwell. 5th edition.

Degand, L. & a.C. Simon (2009). On identifying basic discourse units in speech: theoretical and empirical issues , *Discours*, 4, http://discours.revues.org/5852

Genetti, Carol (2007). Syntax and prosody: Interacting coding systems in Dolakha Newar. *Proceedings of the Thirteenth Annual Meeting of the Southeast Asian Linguistics Societ*y, Iwasaki, Shoichi (ed.). Canberra: Pacific Linguistics.

Gordon, M & P. Ladefoged. 2001. *Phonation types: a cross-linguistic overview*, Journal of Phonetics 29. pp. 383-406.

Harvey, Mark David (2008). *Proto Mirndi: A Discontinuous Language Family in Northern Australia*. Canberra: Pacific Linguistics ACT.

Ladd, D. Robert (2008). *Intonational Phonology*, second edition. Cambridge University Press.

Lehiste, I. (1975). The phonetic structure of paragraphs. In A. Cohen & S.G. Nooteboom (eds), *Structure and Process in Speech Perception*. Springer-Verlag. pp. 195-203.

McGregor, William (2004). *The Languages of the Kimberley, Western Australia.* London, New York: Taylor & Francis.

Mayer, M.(1969). *Frog, where are you?* New York: Dial Books.

Schuetze-Coburn, S. & M. Shapley & E. Weber. (1991). *Units of intonation in discourse: A comparison of acoustic and auditory analyses*. Language and Speech 34. pp. 207-234.

Schultze-Berndt, Eva (2000). *Simple and complex verbs in Jaminjung: A study of event categorisation in an Australian language*. Nijmegen: PhD dissertation, University of Nijmegen.

Swerts, M. & Geluykens. R. (1994). Prosody as a marker of information flow in spoken discourse. *Language and Speech* 37 (1), pp. 21-43. Institute for Perception Research, Eindhoven.

Tseng, Chiu-yu (2010). *Beyond sentence Prosody.* Proceedings INTERSPEECH 2010, pp. 20-29.

Wichmann, A. (2006). Prosody and discourse: a diachronic approach. In: *Proceedings IDP05 Interface Discours-Prosodie (Discourse-Prosody interface*), 8-9 September 2005, Aix-en-Provence, France.

Wichmann. Ann. (2000). *Intonation in Text and Discourse*. London: Pearson Education.

Xu, Yi (2005). Speech Melody as Articulatorily Implemented Communicative Functions. *Speech Communication* 46, pp. 220-251.

La prosodie des marqueurs de thématisation

Tom Velghe

tom.velghe@arts.kuleuven.be
KU Leuven

Abstract

This paper discusses the prosodic properties of PPs introduced by so-called 'thematic markers' (TMCs), such as *en ce qui concerne* ('as for') or *du point de vue de* ('with regard to'). We describe their prosodic properties in terms of the pitch level at their right edge and we show that most TMCs are followed by a relative high pitch level, which means that they end in a major prosodic boundary (Mertens 2008). In addition, we show that TMCs are more often followed by major boundary than initial spatio-temporal adverbials. We provide syntactic and semantic arguments which explain the more frequent strong prosodic boundary at the end of a TMC.

1. Introduction

Une CMT (construction à marqueur de thématisation) est un groupe prépositionnel détaché à gauche de la phrase principale, introduit par une locution prépositive comme *quant à*, *en ce qui concerne*, *pour ce qui est de*, *au niveau (de)*, *en matière de*, *(du/ au) point de vue (de)*, etc. Ces prépositions sont appelées des marqueurs de thématisation parce que les CMT constituent le thème ('theme'; Halliday 1967) de la proposition (cf. Combettes 2003 ; 'marqueurs de topicalisation' − Lagae 2007, 2011a/b ; 'marqueurs thématiques' − Porhiel 2004 ; introducteurs de cadres thématiques'). Leur fonction est d'indiquer *le topique Chinese Style* (Chafe 1976) et/ou *le topique d'à-propos* (Lambrecht 2000) d'une phrase. Dans (1) et (2), la CMT "limite le champ d'application de la proposition centrale à un certain domaine et établit un cadre spatial, temporel ou individuel dans lequel vaut la prédication centrale[1]" (Chafe 1976 : 50). Ainsi dans (1), *au niveau de l'info* indique le domaine où l'assertion est valide. En outre, dans (2), *pour ce qui est de* introduit le topique d'à-propos : l'on affirme quelque chose à propos de Dieudonné.

(1) *Mais quand on peut il faut impérativement regarder BBC News. En ce moment,* au niveau de l'info, *ils sont vraiment au top.* (Corpus Yahoo Answers, De Smet)
(2) Pour ce qui est de Dieudonné, *je ne le défendrai plus.* (Corpus Yahoo Answers, De Smet)

Dans cette contribution nous comparons les CMT avec un autre constituant initial : le circonstant spatio-temporel antéposé (CSTA). Nous les étudions au niveau syntaxique et sémantico-pragmatique et comparons également leurs propriétés prosodiques. Nous montrons que les CMT se terminent plus souvent par une *frontière prosodique majeure* (Mertens 2008) que les CSTA.

2. Propriétés sémantico-pragmatiques et syntaxiques des CMT et CSTA

Les CSTA et les CMT partagent la position initiale, mais nous relevons deux différences importantes, tant au niveau sémantico-pragmatique (i) qu'au niveau syntaxique (ii). Ces différences sous-tendent l'hypothèse d'une frontière syntaxique plus forte à la fin d'une CMT qu'après un CSTA.

2.1. Propriétés sémantico-pragmatiques

Contrairement aux CSTA, les CMT ne limitent ou ne changent pas toujours la signification de la phrase. Dans (3), le

[1] Ma traduction de "a topic Chinese Style limit[s] the applicability of the main predication to a certain restricted domain [...] The topic sets a spatial, temporal, or individual framework within which the main predication holds" (Chafe 1976 : 50).

circonstant spatio-temporel *en ville de Berne* limite la portée de la phrase principale à un lieu précis :

(3) En ville de Berne, *on ne tolérera dorénavant plus les mendiants dans le passage sous la place de la gare.* (Corpus C_Prom, jpa-ch)

Dans (4), la CMT limite elle aussi la portée de la phrase principale. Les parents sont très stricts quand il s'agit de mots vulgaires:

(4) *Mais* en ce qui concerne les mots vulgaires, *oui, ils* [mes parents] *étaient également très stricts.* (Corpus ESLO, 118)

Dans (5) par contre, la CMT *pour ce qui est du plan diplomatique* ne limite pas le champ d'application de la proposition principale. Elle peut parfaitement être supprimée sans changer la signification de la phrase.

(5) Pour ce qui est du plan diplomatique *euh le président Bush a dit qu'il comprenait les objections au projet de résolution franco-américain de l'ONU.* (Corpus C_PROM, jpa)

2.2. Propriétés syntaxiques

Tous les CSTA peuvent être clivés et peuvent se trouver en position finale de la phrase (6). En revanche, les CMT ne peuvent pas toutes être clivées ou apparaître en fin de phrase. Ces transformations sont admises pour les CMT qui limitent le domaine d'application de la phrase principale (7), mais exclues pour celles qui n'affectent pas les conditions de vérité de la proposition (8).

(6) a. En ville de Berne, *les festivités ont commencé.*
b. *C'est* en ville de Berne *que les festivités ont commencé.*
c. *Les festivités ont commencé* en ville de Berne.

(7) a. *Mais* en ce qui concerne les mots vulgaires*, ils étaient très stricts.*
b. *C'est* en ce qui concerne les mots vulgaires *qu'ils étaient très stricts.*
c. *Ils étaient très stricts* en ce qui concerne les mots vulgaires.

(8) a. Pour ce qui est du plan diplomatique *le président Bush a dit qu'il comprenait les objections de l'ONU.*
b. **C'est* pour ce qui est du plan diplomatique *que le président Bush a dit qu'il comprenait les objections de l'ONU.*
c.**Le président Bush a dit qu'il comprenait les objections de l'ONU* pour ce qui est du plan diplomatique.

Les CMT se répartissent donc en deux catégories: (1) d'une part, celles qui peuvent être clivées ainsi que déplacées à la fin de la phrase et qui spécifient la signification de la phrase et (2) d'autre part, celles qui ne peuvent être ni clivées, ni déplacées à la fin de la phrase et qui ne spécifient pas la signification de la phrase.

Les tests syntaxiques et sémantiques de clivage, de déplacement et le fait d'affecter le sens de la proposition centrale sont en général utilisés (e.a. Blanche-Benveniste 1990, Nølke 1990) pour montrer que le constituant en question fait partie de la rection du verbe principal. Les exemples ci-dessus montrent que tous les CSTA sont régis par le verbe, mais que cela ne vaut pas pour l'ensemble des CMT.

3. Cadre descriptif pour l'intonation

Pour l'analyse prosodique, nous adoptons la description de Mertens (2008). La plupart des modèles prosodiques se basent sur la parole lue, constituée d'un nombre limité de phrases relativement courtes, grammaticalement bien formées élaborées à des fins de recherche et prononcées par un groupe de locuteurs dans des circonstances optimales. Le modèle de Mertens, par contre, a été conçu sur la base d'un corpus relativement grand de parole continue, comprenant des

interviews qui font intervenir plusieurs locuteurs.

L'unité prosodique de base dans le modèle de Mertens (2008) est le groupe intonatif (GI) : "une suite d'une ou plusieurs syllabes dont la dernière syllabe pleine porte un accent final (AF). Par syllabe pleine on entend toute syllabe qui comporte une voyelle autre que le schwa" (Mertens 2008 : 94). Outre les syllabes atones (NA) et un AF, le GI peut également comporter un accent initial (AI) facultatif ce qui donne la structure interne suivante, où les crochets entourent des parties facultatives :

GI = [[NA] AI] [NA] AF

D'après Mertens (2008 : 93), chaque GI entraîne une frontière prosodique d'un certain degré qui dépend de la hauteur relative de la syllabe accentuée finale (voir aussi Rossi 1999, Martin 1975/1978). Ainsi dans la figure 1, les syllabes accentuées finales représentées par les traits en gras de a), b) et c) sont séparées des syllabes précédentes par un grand intervalle mélodique dont la taille varie avec la tessiture du locuteur, mais est le plus souvent de l'ordre de 4 demi-tons. Le niveau de hauteur à la fin du GI dans a) correspond à l'infra-bas (B-B-), qui correspond au plancher de la tessiture. Dans c), la syllabe finale du GI comporte une montée intra-syllabique qui part du niveau haut (H/H). Quant à d), l'intervalle entre les syllabes tonique et prétonique est moins de 4 demi-tons et Mertens (2008) le considère donc comme un intervalle mineur : /BB.

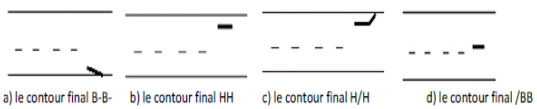


Figure 1. Quelques exemples de contours finals.

Les contours intonatifs qui peuvent apparaître à la fin des GI sont très variés. Mertens (2008) considère les contours BB, \BB, /BB, B/B et B\B comme des frontières mineures et les contours B-B-, BH, HB, H/H, \HH, /HH, H/H et H\H comme des frontières majeures.

L'auteur avance l'hypothèse que dans certaines constructions syntaxiques comme la dislocation à gauche, le clivage et certains adverbes de phrase, une frontière prosodique majeure apparaît à la charnière entre la partie antéposée (l'élément disloqué, le foyer de la clivée ou l'adverbe de phrase) et le noyau verbal.

Pour la représentation graphique de la prosodie, nous faisons appel à Prosogram (Mertens 2004) qui affiche la hauteur mélodique telle qu'elle est perçue par l'auditeur moyen. L'auteur part de l'observation que les variations mélodiques sur les noyaux vocaliques sont déterminantes du point de vue perceptif (House 1990 ; d'Alessandro & Mertens 1995). Cela implique que pour obtenir la stylisation tonale, l'on ne tient pas compte de toutes les variations du signal acoustique puisqu'elles ne sont pas toutes audibles (e.a. d'Alessandro & Mertens 1995). Pour que le locuteur puisse les percevoir, elles doivent être d'une ampleur et durée suffisantes.

La figure 2 illustre la représentation acoustique obtenue à l'aide de Prosogram pour l'énoncé *Mais cette journée de guerre n'a pas empêché les ministres des affaires étrangères de la ligue arabe de se réunir dans la capitale libanaise* (Corpus C_Prom). La fréquence fondamentale (F_0) est représentée par la courbe en bleu sur une échelle en demi-tons : les lignes pointillées horizontales indiquent une calibration de l'axe vertical qui représente la hauteur mélodique; la distance entre deux lignes successives est de 2 demi-tons. La ligne en vert donne l'intensité en dB et la ligne en zigzag indique les parties voisées. Le trait noir épais donne une estimation de la hauteur perçue par un auditeur moyen. Cette valeur stylisée est utilisée pour interpréter l'intonation.

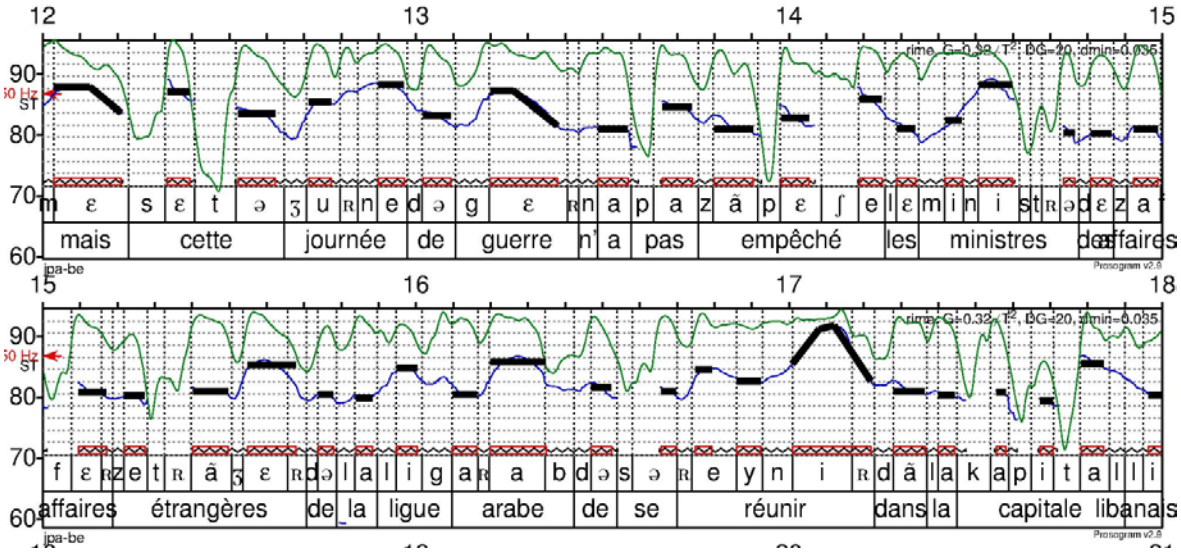


Figure 2. Contour mélodique stylisé (Prosogram) de l'énoncé *Mais cette journée de guerre n'a pas empêché les ministres des affaires étrangères de la ligue arabe de se réunir dans la capitale libanaise.* (Corpus C_PROM, jpa-be)

(Mais) (cette journée) (de guerre) (n'a pas) (empêché) (les ministres)
AF AF AF AF AF AF
HB h.....b HH bb HB bb /BB \b....b HH b.....b H b...
(des affaires étrangères) (de la ligue) (arabe) (de se réunir) (dans la capitale)
AF AF AF AF AF
.....b HH b....b HH bb HH b.....b BH b.....bHH
(libanaise)
AF
.....b HH

Figure 3. Transcription du fragment 2.

En appliquant la description de Mertens (2008) à l'extrait de la figure 2, nous obtenons la transcription suivante. À la première ligne, les groupes intonatifs sont délimités par des parenthèses, avec à la deuxième ligne, la mention des syllabes sur lesquelles tombent les accents finals (AF). Les contours mélodiques des différentes syllabes[2] sont indiqués à la troisième ligne.

4. Les corpus utilisés

Les données analysées proviennent de trois corpus : Rhapsodie[3], C_Prom[4] et ESLO[5]. Les fichiers exploités pour l'analyse des CMT ont une durée totale d'environ 31 heures, avec ESLO comme corpus le plus grand (environ 27 heures et demie). Pour les CSTA, un échantillon plus petit d'une heure et demie a été analysé. Cet échantillon contient seulement des fichiers de Rhapsodie et de C_Prom. Pour C_Prom et Rhapsodie, un alignement au niveau de la syllabe était déjà

[2] Les majuscules doubles représentent les AF et les minuscules les syllabes atones (une seule majuscule indique l'AI).

[3] http://www.projet-rhapsodie.fr/
[4] https://sites.google.com/site/corpusprom/
[5] http: //www.univ-orleans.fr/eslo/

disponible. Pour les fragments d'ESLO par contre nous avons ajouté l'alignement à la main pour les exemples étudiés.

5. Résultats

5.1. La prosodie des marqueurs de thématisation

Au total, 55 occurrences de CMT ont été repérées dans notre corpus. Dans le premier tableau, nous énumérons les contours relevés à la fin des CMT. Il apparaît que 73% des CMT se terminent par une frontière prosodique majeure, avec le contour HH comme contour le plus fréquent. Par contre, 27% des CMT dans notre corpus se terminent par une frontière prosodique mineure.

Type de frontière prosodique	***Fréquence***
Frontière prosodique majeure	72,8% (40)
HH	43,6% (24)
H/H	12,7% (7)
H\H	5,5% (3)
HB	5,5% (3)
BH	5,5% (3)
Frontière prosodique mineure	27,2% (15)
BB	18,2% (10)
/BB	7,3% (4)
B\B	1,7% (1)
Total	*100% (55)*

Tableau 1: les frontières prosodiques à la fin des CMT

Dans (9), la CMT est accompagnée d'une frontière prosodique majeure. Elle est constituée de plusieurs GI dont le dernier se termine par le contour HH.

(9) *(Pour ce qui est)$_{HH}$ (de notre représentation)$_{BB}$ (de l'intona**tion**)$_{**HH**}$ (en fait)$_{HH}$ (on reprend)$_{/BB}$ (le terme)$_{BB}$ (de profil mélodique)$_{HH}$* (Corpus C_Prom, cnf-fr)

Dans (10) par contre, la CMT se termine par une frontière prosodique mineure, à savoir, le contour BB.

(10) *(En ce qui concerne)$_{HB}$ (l'hydrau**lique)**$_{**BB**}$ (je vois par exemple)$_{HH}$ (euh en Charente)$_{BH}$* (Corpus ESLO, fichier 012)

5.2. La prosodie des CSTA

Des tableaux 1 et 2 il ressort clairement que la prosodie des CSTA est différente de celle des CMT. 73% des CMT sont accompagnées d'une frontière prosodique majeure, contre seulement 40% pour les CSTA. La frontière mineure est donc plus fréquente à la fin d'un CSTA qu'à la fin d'une CMT. Le tableau 2 montre aussi que le contour BB est le plus fréquent pour clôturer un CSTA.

Type de frontière prosodique	**Fréquence**
Frontière prosodique mineure	
BB	38,5% (44)
/BB	10,4% (12)
B/B	8,7% (10)
\BB	1,7% (2)
	Total : 59,1% (68)
Frontière prosodique majeure	
HH	15,7% (18)
H/H	6,1 % (7)
HB	1,7% (2)
BH	12,2 % (14)
\HH	0,9% (1)
/HH	3,4 % (4)
B-B-	0,9% (1)
	Total : 40,9% (47)
	Total : 100% (115)

Tableau 2: les frontières prosodiques à la fin des CSTA

Dans (11), le CSTA *en ville de Berne* se termine par une frontière prosodique mineure (BB).

(11) *(en ville$_{H}$ de **Berne)**$_{**BB**}$ (on ne to$_{H}$lérera)$_{BB}$ (dorénavant plus)$_{HH}$ (les mendiants)$_{/BB}$ (dans le passage)$_{BB}$ (sous la place de la gare)$_{/HH}$ (et aux abords des accès à ce passage)$_{HH}$* (Corpus C_PROM, jpa-ch)

De ce qui précède, nous pouvons conclure que dans la majorité des cas les frontières prosodiques à la fin des CMT sont plus fortes que celles à la fin des CSTA.

6. Conclusion et perspectives

Malgré leur position initiale commune, les CMT et les CSTA présentent des propriétés sémantico-pragmatiques, syntaxiques et prosodiques différentes. En premier lieu, tous les CSTA limitent la portée de la phrase à un certain domaine, alors que ce n'est pas nécessairement le cas pour les CMT. En deuxième lieu, tous les CSTA peuvent être clivés ou déplacés à la fin de la phrase sans changer le sens de la phrase. Pour les CMT ces mêmes transformations peuvent entraîner un changement de sens. Au niveau prosodique, les CMT sont plus souvent suivies d'une frontière prosodique majeure que les CSTA.

Ces trois différences sous-tendent l'hypothèse d'une frontière syntaxique plus forte à la fin d'une CMT qu'après un CSTA et semblent suggérer que la CMT entretient une relation moins étroite avec le noyau verbal (Blanche-Benveniste 1990, Nølke 1990).

Dans de recherches futures, il serait intéressant de dépouiller des corpus plus grands pour examiner si la présence d'une frontière majeure est influencée par d'autres facteurs, tels que le nombre de syllabes dans le constituant antéposé, le débit et la tessiture du locuteur ou le type de parole.

Références

d'Alessandro, Ch. & P. Mertens. (1995). Automatic pitch contour stylization using a model of tonal perception. *Computer Speech and Language* 9:3, pp. 257-288.

Avanzi, M., A.C. Simon, J.P. Goldman & A. Auchlin. (2010). C-PROM, Un corpus de français parlé annoté pour l'étude des proéminences. *Actes des 23èmes journées d'étude sur la parole* (Mons, Belgique, 25-28 mai 2010).

Blanche-Benveniste, Cl. et al. (1990). *Le français parlé* : *Études grammaticales*. Paris : CNRS.

Branca-Rosoff S., S. Fleury, Fl. Lefeuvre & M. Pires. (2012). *Discours sur la ville, Corpus de Français Parlé Parisien des années 2000* (CFPP2000).

Chafe, W. (1976). Givenness, contrastiveness, definiteness, subjects, topics and point of view. Charles N. Lie (ed.), *Subject and topic*. New York: Academic Press.

Combettes, B. (2003). Les marqueurs de topicalisation en français : un cas de grammaticalisation. Combettes, B., S. Schnedecker & A. Theissen, (eds), *Ordre et distinction dans la langue et le discours*. Paris: Champion, pp. 149-167.

Halliday, M.A.K. (1967). Notes on transitivity and theme in English, part II. *Journal of Linguistics* 3, pp. 199-244.

House, D. (1990). *Tonal Perception in Speech, Lund*: Lund University Press.

Lagae, V. (2007). Left-detachment and topic-marking in French: the case of quant à and en fait de. *Folia linguistica* 41, pp. 327-355.

Lagae, V. (2011a). À propos de: un marqueur thématique très particulier. Amiot, D., W. De Mulder, Moline, E. & Stosic, D. (eds.), *Ars Grammatica. Hommages à Nelly Flaux.* Berne: Peter Lang, pp. 273-288.

Lagae, V. (2011b). Le paradigme des marqueurs thématiques en français: essai de typologie. E. Comes & Miculescu, S. (eds.), *La construction d'un paradigme - Actes du XVIIIe Séminaire de Didactique Universitaire Constanta 2010.* Cluj: Editura Echinox, pp. 53-74.

Lambrecht, K. (2000). Information Structure and Sentence form: topic, focus and the mental representations discourse referents. Cambridge: Cambridge University Press.

Martin, Ph. (1975). Analyse phonologique de la phrase française. *Linguistics* 146, pp. 35-68.

Martin, Ph. (1978). L'intonation de phrases à structure non connexe. *Bulletin de l'Institut de Phonétique* ULB 12:1, pp. 97-106.

Mertens, P. (2004). The Prosogram : Semi-Automatic Transcription of Prosody based on a Tonal Perception Model. B. Bel & I. Marlien (eds.), *Proceedings of Speech Prosody 2004*. Nara, 23-26 March.

Mertens, P. (2008). Syntaxe, prosodie et structure informationnelle: une approche prédictive pour l'analyse de l'intonation dans le discours. *Travaux de Linguistique* 56:1, pp. 87-124.

Nølke, H. (1990). Les adverbiaux contextuels: problèmes de classification. *Langue française* 88, pp. 12- 27.

Porhiel, S. (2004). Les introducteurs de cadre thématique. *Cahiers Lexicologiques* 85, pp. 9-45.

Rossi, M. (1999). *L'intonation, le système du français: description et modélisation*. Paris: Ophrys.